# Advantages to and Challenges of Using Ratings of Observed Teacher-Child Interactions

A Dissertation

Presented to

The Faculty of the Curry School of Education

University of Virginia

In Partial Fulfillment

Of the Requirements for the Degree

Doctor of Philosophy

By

Anne E. Henry

August 2010

# Abstract

Robert C. Pianta, Advisor

This dissertation presents three independent studies that are linked in their focus on issues relevant to observing teacher-child interactions. The first study confirms the degree to which observed teacher-child interactions predict children's academic and social outcomes, and the second two studies inform the challenges in maintaining this predictive validity when using observations in large-scale contexts. Study 1 illustrates how children exposed to high quality teacher-child interactions consistent with effective classroom organization and instructional support in both pre-kindergarten and kindergarten scored significantly higher on assessments of their language and literacy skills. Study 2 examines the relationship between observed scores using the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, and Hamre, 2008) and observation protocol characteristics, specifically the day of the week, month of the year, and duration of an observation cycle in minutes. CLASS scores were stable across these factors, with a few exceptions for the CLASS domain of Classroom Organization. Study 3 describes the extent of rater calibration resulting from a large-scale training effort by the Office of Head Start and explores rater characteristics that predict calibration. The majority of raters trained in this large-scale effort passed an initial calibration assessment and rater beliefs predicted the degree of calibration. Collectively, these three studies demonstrate that observational assessment can provide meaningful information about teachers and children in large-scale contexts despite challenges faced in planning and implementation.

Department of Leadership, Foundations, and Policy

Curry School of Education

University of Virginia

APPROVAL OF THE DISSERTATION

This dissertation, "Advantages of and Challenges to Using Ratings of Observed Teacher-Child Interactions," has been approved by the Graduate Faculty of the Curry School of Education in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

________________________________________
Dr. Robert Pianta, Advisor

________________________________________
Dr. Jason Downer

________________________________________
Dr. Patrick Meyer

________________________________________
Dr. James Wyckoff

6/24/10 Date

# DEDICATION

This work is dedicated to Micah Cash, for his unquestioning patience, support, and sacrifice, that I might achieve this mutual goal.

# ACKNOWLEDGEMENTS

I would like to express my deepest gratitude for the mentors, colleagues, and friends who have contributed to my development as a scientist. Thank you to the faculty on my dissertation committee and in the Educational Psychology-Applied Development Science program, both within and outside of the Curry School of Education, for laying the foundation through the quality of your coursework, special sessions, and feedback. I extend a special word of appreciation to my advisor, Bob Pianta, for the energy, wisdom, and generosity that is evident in his mentoring. I admire his passion for his work, grounded in motivation to improve experiences for teachers and children. Through him, I have learned the importance of being able to tell a story in varying levels of detail, while keeping the big picture in mind. I thank Jason Downer and Bridget Hamre, for being responsive and clear when I need them, but also for giving me courage and room to grow. I thank my many colleagues, past and present, at the Center for Advanced Study of Teaching and Learning, particularly Jennifer Locasale-Crouch, Tim Curby, Amy Luckner, Claire Cameron-Ponitz, and Ginny Vitiello, whose conversation makes this work so engaging and fun. I am also appreciative of Leslie Booren, whose friendship bridges those times when it is not. Finally, I must acknowledge the generous financial support throughout my doctoral studies from the Institute of Education Sciences, U.S. Department of Education. This work was funded through Grant #R305B040049 to the University of Virginia.

## TABLE OF CONTENTS

1. DEDICATION……………………………………………………………………iv
2. ACKNOWLEDGEMENTS……………………………………….…………………v
3. LIST OF TABLES………………………………………………………….....vi
4. LIST OF FIGURES………………………………………………………...vii
5. Conceptual Link: Advantages to and Challenges of Using Ratings of Observed Teacher-Child Interactions: Rationale and Conceptual Links Across the Three Dissertation Studies……….………………………………………………………..1
6. Study 1: Power of Two: Impact of Two Years of High Quality Teacher-Child Interactions………………..…………………………………………………...2
7. Study 2: Accounting for Variance in an Observational Measure of Teacher Quality…………………………………………………………..……………..3
8. Study 3: Rater Calibration When Observational Assessment Occurs at Large-Scale: Degree of Calibration and Characteristics of Raters Associated with Calibration………………………………………………………………………..4

# LIST OF TABLES

TABLE Page

1. *Study 1:* CLASS Cutoff (Mean) Scores for Group Assignment……………………..68
2. *Study 1:* Assessing Bias Reduction and Group Comparability…………………........69
3. *Study 1:* Effects of Teacher-Child Interactions on Kindergarten Spring Outcomes…70
4. *Study 2:* Frequency of Videotapes Recorded on a Given Day or in a Given Month………………………………………………………………………….105
5. *Study 2:* Fixed Effects Estimates and Variance-Covariance Estimates for CLASS Domains……………………………………………………………………….106
6. *Study 2:* Fixed Effects Estimates and Variance-Covariance Estimates for CLASS Dimensions..……………………………………………………………………107
7. *Study 3:* Descriptive Statistics for Adjacent Calibration Scores, Percent Within-1 of Master-Code……………………………………………………………………..160
8. *Study 3:* Descriptive Statistics for Valence Distance Scores……………………….161
9. *Study 3:* Descriptive Statistics for Absolute Distance Scores………………………162
10. *Study 3:* Fixed Effects Estimates and Variance-Covariance Estimates for Models of the Predictors of Absolute Distance, Full Sample……………………………..163-166
11. *Study 3:* Fixed Effects Estimates and Variance-Covariance Estimates for Models of the Predictors of Absolute Distance, Survey Sample…………………………167-177

# LIST OF FIGURES

FIGURE Page

1. *Study 1:* Overall Design of Data Analysis for Each of the CLASS Domains………64

2. *Study 1:*  Matching Process………………………………………………………….65

3. *Study 1:* Propensity Score Distribution for Each CLASS Domain, Pre- and Post-Match……………………………………………………………………………...66

# Advantages to and Challenges of Using Ratings of Observed Teacher-Child Interactions:

# Rationale and Conceptual Links across the Three Dissertation Studies

Anne E. Henry

## The Three-Manuscript Dissertation: Overview

This dissertation presents a line of research exploring the advantages and challenges of using ratings of observed teacher-child interactions to assess and improve teacher quality. The dissertation is written according to guidelines in the Curry School of Education’s Dissertation Manual: Guidelines for Doctoral Dissertations for the manuscript-style dissertation option.

The Curry School Guidelines require the student to take a lead role on two research papers, contribute to a third research paper, and submit an additional document that articulates the conceptual link among the manuscripts. I am the lead author on all three of the studies described here, and they will be submitted to refereed journals upon completion. The remainder of this document covers the rationale for the program of research and conceptual links among the studies, and each of the three manuscripts presented in full.

## Advantages to and Challenges of Using Ratings of Observed Teacher-Child Interactions

Teachers play an important role in children’s development, and at least in terms of achievement outcomes, differences between teachers account for more variation in children’s skills than do differences between schools (Nye, Konstantopoulos, & Hedges, 2004; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004)(Nye, K, Hedges, 2004; Rivkin 2005; Rockoff 2004). There has been increasing pressure in recent years to identify the characteristics of teachers who are most effective in improving children’s outcomes and to improve the quality of teachers who are less effective. For kindergarten-12th grade settings, this pressure is associated with the authorization (and reauthorization) of No Child Left Behind (NCLB) and related changes to teacher licensing systems at the state level. In early childhood education, pressure to identify effective teachers is evident in the development of state-level Quality Rating Systems (QRS; Barnett, Epstein, Friedman, Boyd, & Hustedt, 2008; National Child Care Information and Technical Assistance Center, 2007, 2009). Suddenly, policymakers and school administrators are charged not only with measuring and improving teacher quality, but doing so at a large scale.

### Classroom Observation as a Tool for Assessing and Improving Quality

Classroom observation is increasingly a component of systems to assess and improve quality at large-scale. For example, the Improving Head Start for School Readiness Act of 2007 required the Office of Head Start to include a valid and reliable observational tool for assessing classroom quality in grantee monitoring reviews (U.S. Department of Health and Human Services, Administration for Children and Families, & Office of Head Start, 2008). Also, a portion of $22 million in grants designated for research on teacher effectiveness by the Bill and Melinda Gates Foundation will be used to videotape and observe teachers (Medina, 2009; Robelen, 2008, 2009). Some of the best video examples will be posted online as a resource for teachers and students (Gates, 2009).

Perhaps the most extensive use of classroom observation to date has been in the context of Quality Rating Systems (QRS) for early learning settings. QRS typically involve five elements – program standards, accountability measures, program and practitioner outreach and support, financial incentives, and consumer education (National Association of Child Care Resource and Referral Agencies, 2009; National Child Care Information and Technical Assistance Center, 2007). Classroom observation can be used within QRS as a measure of how well programs meet standards or as a tool for improvement through the application of environment rating scales. In fact, the National Association of Child Care Resource and Referral Agencies (NACCRRA) recommends that states developing or implementing QRS incorporate annual rating visits that include on-site observation (National Association of Child Care Resource and Referral Agencies, 2009). In 2009, 12 states were using environment rating scales as part of their assessment of standards, two more were using them for program improvement, and one state was using them as an alternative pathway for meeting quality standards (National Association of Child Care Resource and Referral Agencies, 2009).

A series of scales developed by the Frank Porter Graham Child Development Institute at the University of North Carolina have frequently been used to observe learning environments and program quality for these purposes (Bryant, 2000; National Association of Child Care Resource and Referral Agencies, 2009; National Child Care Information and Technical Assistance Center, 2007; Norris, Dunn, & Eckert, 2003; Tout, Starr, & Cleveland, 2008), including the Early Childhood Environment Rating Scale – Revised (ECERS-R; Harms, Clifford, & Cryer, 1998), Infant/Toddler Environment Rating Scale-Revised (ITERS-R; Harms, Cryer, & Clifford, 1990), School-Age Care Environment Rating Scale (SACERS; Harms, Jacobs, & White, 1996), and the Family Day Care Rating Scale (FDCRS; Harms & Clifford, 1989). These scales measure aspects of learning settings such as daily routines, activities and materials in the classroom, and interactions between teachers and children. Ratings on the ECERS-R and ITERS-R appear consistent with ratings on some states' QRS; higher ratings on the ECERS were associated with higher "star" ratings on QRS in Oklahoma (Norris et al., 2003) and North Carolina (Bryant, 2000). Improvement in program quality according to the ECERS-R has also been documented (Norris et al., 2003). Two states, Minnesota and Virginia, have been piloting the use of another observational tool, the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008a) in their QRS (Tout, Zaslow, Halle, & Forry, 2009).

### Associations between Observed Scores and Children's Outcomes

In QRS, data from multiple measures of program quality are typically combined to create a composite quality score that is tied to each program's rating. Much research is still needed to determine the association between composite scores from QRS and children's outcomes, though a study on Qualistar, a QRS in Colorado, indicates that few significant relationships between ratings and outcomes exist (Zellman & Perlman, 2008). There are a couple of possible explanations for such weak or nonexistent relationships, including lack of clarity in how to best create composite scores, or individual indicators being poorly measured or only weakly linked to children's outcomes (National

Association of Child Care Resource and Referral Agencies, 2009; Tout et al., 2009; Zellman & Perlman, 2008; Zellman, Perlman, Le, & Setodji, 2008). For example, teachers' level of education is frequently measured as one component of a QRS (National Child Care Information and Technical Assistance Center, 2007), but is not consistently associated with children's academic or social outcomes across seven large-scale studies (Early et al., 2007).

One possible way to improve the power of QRS composite scores to predict children's outcomes is to rely more heavily on indicators that have been shown to be related to children's outcomes. For example, observations of classroom environments are also common components of QRS (National Association of Child Care Resource and Referral Agencies, 2009; National Child Care Information and Technical Assistance Center, 2007; Tout et al., 2009) and are associated with QRS composite scores in at least two states (Bryant, 2000; Norris et al., 2003). More importantly, observations of classroom environments are predictive of children's outcomes (e.g. Burchinal et al., 2008; Burchinal, Vandergrift, Pianta, & Mashburn, 2009; Connor, Son, Hindman, & Morrison, 2005; Curby, LoCasale-Crouch et al., 2009; Hamre & Pianta, 2005; Howes et al., 2008; Mashburn et al., 2008; Pianta, La Paro, Payne, Cox, & Bradley, 2002; Rimm-Kaufman, Curby, Grimm, Nathanson, & Brock, 2009). In fact, these measures of children's direct experiences in classrooms appear more predictive of their outcomes than measures of the structural features of classrooms such as teacher qualifications or the location of the program (Howes et al., 2008).

More specifically, ratings from observations focused on the interactions between teachers and children are predictive of children's academic and social outcomes (e.g.

Burchinal et al., 2008; Burchinal et al., 2009; Connor et al., 2005; Curby, LoCasale-Crouch et al., 2009; Hamre & Pianta, 2005; Howes et al., 2008; Mashburn et al., 2008; Pianta et al., 2002; Rimm-Kaufman et al., 2009). For example, children in classrooms rated higher quality in terms of observed interactions with teachers who are respectful and responsive to their needs demonstrated greater levels of social competence (Burchinal et al., 2009; Mashburn et al., 2008), greater growth in phonological awareness (Curby, LoCasale-Crouch et al., 2009), and higher vocabulary and decoding scores (Connor et al., 2005). Children in classrooms that received higher scores on observational scales of the instructional support provided by teachers demonstrated higher levels of academic and language skills (Burchinal et al., 2008; Howes et al., 2008; Mashburn et al., 2008). Children whose teachers were observed to have established clear routines and proactive approaches to discipline have greater levels of behavioral and cognitive self-control, spend less time off-task (Rimm-Kaufman et al., 2009), and greater gains in math skills (Curby, LoCasale-Crouch et al., 2009).

Moreover, higher scores on observational measures of emotionally and instructionally supportive teacher-child interactions appear to moderate the relationship between children’s level of risk and outcomes (Burchinal et al., 2009; Hamre & Pianta, 2005). Children whose mothers had less than a 4-year college degree, who also had teachers who provided moderate to high levels of instructional support, demonstrated levels of achievement similar to their low-risk peers at the end of the first grade year (Hamre & Pianta, 2005). Similarly, children identified as displaying early behavioral, social, or academic problems at the end of kindergarten demonstrated levels of achievement similar to their low-risk peers if they were exposed to first grade classrooms high in emotional support (Hamre & Pianta, 2005). For children who are poor, the relationship between the level of emotional support they experience from their teachers and their level of social and behavioral competence is strongest at the highest levels of emotional support (Burchinal et al., 2009).

## Challenges to Establishing Best Practices in Observation Methodology

Given the evidence described above that ratings from observational measures of teacher-child interactions can predict children's academic, social, and behavioral competencies, there is ongoing interest in the use of such scales for assessing teacher quality, particularly in the context of state QRS or large randomized control trials. Recent emphasis on large-scale observation calls for attention to best practice in observational assessment. Decisions that evaluators and researchers make in planning an observation protocol impact the reliability and validity of the data as well as the cost of collection.

One major challenge to using observational tools is confirming their reliability and validity. The interactions among teachers and children in classroom environments are complex, and scores from observational measures of these interactions are subject to multiple sources of variation. There is evidence that a greater proportion of variation in teacher quality is accounted for between teachers than between schools (Nye et al., 2004). Variation in quality between teachers could be explained by characteristics of the children they work with or the teachers themselves. For example, children's developmental skills at the beginning of the school year or teachers' level of experience with a certain grade level may impact performance.

Importantly, there is also variation in scores that occurs within teachers who are observed across multiple occasions (Curby, Brock, & Hamre, 2009; J. P. Meyer, Henry, & Mashburn, 2009; Pianta, La Paro et al., 2008a; Zellman & Perlman, 2008). Observation protocols often call for multiple observations of the same teacher, across multiple days during the school year and/or across multiple observation occasions within the same day. Differences in scores from one observation occasion to the next may be providing systematic information about children, teachers, and classrooms. For example, scores could vary due to activity settings (i.e. math, reading, transitions), group size (i.e. large group, small group), or time of day observed (i.e. morning, after lunch). Perhaps children are less focused on instruction following recess, or more likely to interact with their peers in small group settings. Alternatively, variation within teachers observed on multiple occasions could be explained by characteristics of the observation protocol, such as rater or the duration of the observation occasion. Researchers and evaluators in charge of designing observation protocols must consider potential sources of variation such as these and determine which characteristics of people, settings, and protocol can or should be controlled for during observation or later analyses. There are often multiple goals for observation (i.e. research, quality improvement, consumer awareness) and these should be taken into account when making protocol decisions as well (Tout et al., 2009).

Tools for assessing the reliability and validity of observational data have historically been adapted from those developed for assessing the psychometric properties of individual-level data (Hintze, 2005; Raudenbush & Sampson, 1999). Reflecting on the need for a different approach to evaluate the properties of settings-level measurement, Raudenbush and Sampson (1999) coined the term "ecometric assessment" (p. 3) and incorporated three analytic strategies in their example of observing Chicago neighborhoods: 1) item response modeling, 2) generalizability theory, and 3) factor analysis.

Of these analytic approaches, the one that has been used to understand multiple sources of variation present in observed scores is multivariate generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Generalizability theory is a set of strategies that can be used to evaluate the degree to which a given set of observations generalizes to a more extensive set of observations of that individual or setting. Evaluators can identify portions of variation that can be accounted for by situational variables such as time of day or rater. Related to generalizability theory, decision theory can help evaluators take advantage of these variance estimates in designing protocols (Brennan, 2001).

Although theories such as generalizability theory are available to account for and inform variation in observed scores, they are not always used to inform very practical decisions in planning observation protocols. The logistics of classroom observation can be challenging and expensive, particularly when observation occurs at a large scale. When finite resources are at stake, should more teachers be observed, or fewer teachers observed but across multiple occasions? Is it better to observe more often in the fall or the spring of the year, early in the school-day or late in the afternoon? Research could tell us which options are better for the reliability and validity of the data, but more often decisions are based on availability of schools and teachers. States currently have different rules concerning sampling for QRSs (Tout et al., 2009). Often teachers, even new teachers, are only observed on one occasion if at all (National Council on Teacher

Quality, 2007), even though we have reasons to believe that more observations are necessary to calculate reliable estimates of quality (Hintze & Matthews, 2004).

Even if an original protocol is based in research, observation protocol can sometimes shift the course of a project, either in response to feedback from participants or in response to change in available resources. For example, observers for Minnesota's QRS switched the order of observational tools used in response to programs' level of comfort and familiarity in being observed (Tout et al., 2008). More research is needed to identify the impact of these types of decisions on observed scores.

Another challenge on the logistical side of observational assessment is the decision of who should be hired to collect observational data. Training large groups of observers to assess classrooms can be time-consuming and expensive. In North Carolina, individual programs are only assessed every three years, yet there is still a waiting list for assessments because there are too few trained observers (Zellman & Perlman, 2008). QRS often assign both rating and coaching responsibilities to the same individuals to cut costs, but this practice raises concerns regarding reliability (Tout et al., 2009; Zellman & Perlman, 2008). Observers who are also quality improvement personnel may find it difficult to rate classrooms objectively once they have established relationships with program staff or feel responsible for the success of the coaching experience (Tout et al., 2008; Zellman & Perlman, 2008).

Observer reliability is crucial to the success of large-scale observation efforts. This is particularly true when evaluators are trying to determine if quality has changed over time or to compare the quality of individual programs with each other (Zaslow, Tout, Halle, & Forry, 2009). More work is needed to identify what training and support mechanisms must be in place to ensure and track observer reliability. Currently, there is a lot of variation in these mechanisms, and reliability is sometimes tracked at the state level and sometimes tracked at the local level (National Association of Child Care Resource and Referral Agencies, 2009). Finally, research is also needed to identify background characteristics of observers that are predictive of their ability to assign reliable classroom ratings once trained. Such research would have implications for hiring and retention of observers in large-scale contexts.

## A Three-Study Approach

The aim of this dissertation is to present a line of research which further confirms the power of observed teacher-child interactions to predict children's academic and social outcomes, and given effects, informs common challenges in maintaining this predictive validity when using observations to assess teacher-child interactions in large-scale contexts. These studies complement each other because they have implications for handling potential correlates of observed teacher-child interaction quality at the point of study design as well as during data analysis.

In the first study, I examine the impact of high quality teacher-child interactions in pre-kindergarten and kindergarten, as measured by the Classroom Assessment Scoring System (CLASS; Pianta, La Paro et al., 2008a), on children's outcomes at the end of kindergarten, using propensity score matching to carefully control for children's selection into high quality early learning settings. This study further establishes the predictive validity of this observational tool for measuring teacher-child interactions. Understanding the power of selection bias in studies of the effects of teacher quality is important when planning observational studies as well as in analyzing data. Study 2 examines the stability of CLASS scores given differences in the month, day, or duration of observations by evaluating the degree of within-teacher variation in scores that is accounted for by these factors. To the extent that these protocol characteristics predict the level of quality observed, they should be controlled for in study design or later analyses. Study 3 evaluates the extent of rater calibration when raters are trained in a large-scale context and also explores the association between potential raters' background characteristics and their level of calibration following initial training to use the CLASS. This information could be used by evaluators who need to hire and train large workforces of observers for large-scale studies.

## References

Barnett, W. S., Epstein, D. J., Friedman, A. H., Boyd, J. S., & Hustedt, J. T. (2008). *The state of preschool 2008: State preschool yearbook*. New Brunswick, NJ: National Institute for Early Education Research, Rutgers University.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Bryant, D. M. (2000). *Validating North Carolina's 5-star child care licensing system*. Chapel Hill, NC: Frank Porter Graham Child Development Center.

Burchinal, M., Howes, C., Pianta, R. C., Bryant, D., Early, D. M., Clifford, R. M., et al. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher–child interactions and instruction. *Applied Developmental Science, 12*(3), 140-153.

Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2009). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly*.

Connor, C. M. D., Son, S. H., Hindman, A. H., & Morrison, F. J. (2005). Teacher qualifications, classroom practices, family characteristics, and preschool experience: Complex effects on first graders' vocabulary and early reading outcomes. *Journal of School Psychology*, *43*(4), 343-375.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Curby, T. W., Brock, L. L., & Hamre, B. K. (2009). The role of consistency in preschool teacher-child interactions. *Manuscript under review*.

Curby, T. W., LoCasale-Crouch, J., Konold, T. R., Pianta, R. C., Howes, C., Burchinal, M., et al. (2009). The Relations of Observed Pre-K Classroom Quality Profiles to Children's Achievement and Social Competence. *Early Education & Development*, *20*(2), 346-372.

Early, D. M., Maxwell, K. L., Burchinal, M., Alva, S., Bender, R. H., Bryant, D., et al. (2007). Teachers' education, classroom quality, and young children's academic skills: Results from seven studies of preschool programs. *Child Development, 78*(2), 558-580.

Gates, B. (2009). *2009 annual letter*: Bill and Melinda Gates Foundation. http://www.gatesfoundation.org/annual-letter/Pages/2009-bill-gates-annual-letter.aspx

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*(5), 949-967.

Harms, T., & Clifford, R. (1989). *The Family Day Care Rating Scale*. New York: Teachers College Press.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *The Early Childhood Environment Rating Scale: Revised edition*. New York: Teachers College Press.

Harms, T., Cryer, D., & Clifford, R. M. (1990). *Infant/Toddler Environment Rating Scale*. New York: Teachers College Press.

Harms, T., Jacobs, E., & White, D. (1996). *School Age Care Environment Rating Scale*. New York: Teachers College Press.

Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review, 34*(4), 507.

Hintze, J. M., & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review, 33*(2), 258-271.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., et al. (2008). Ready to learn? Children's pre-academic achievement in pre-Kindergarten programs. *Early Childhood Research Quarterly, 23*(1), 27-50.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O., Bryant, D., et al. (2008). Measures of classroom quality in prekindergarten and children’s development of academic, language, and social skills. *Child Development, 79*(3), 732-749.

Medina, J. (2009, September 1). A 2-year study to learn what makes teachers good. *The New York Times*. Retrieved from http://cityroom.blogs.nytimes.com/2009/09/01/a-2-year-study-to-learn-what-makes-teachers-good/

Meyer, J. P., Henry, A. E., & Mashburn, A. J. (2009). *The influence of occasion on the reliability of classroom observations: An application of multivariate generalizability theory*. Paper presented at the Northeastern Educational Research Association, Rocky Mount, CT.

National Association of Child Care Resource and Referral Agencies. (2009). *Comparison of Quality Rating and Improvement Systems (QRIS) with Department of Defense standards for quality* (No. 724-0714). Retrieved from http://www.naccrra.org/publications/naccrra-publications/comparison-qris-dod-quality-standards-2009.php

National Child Care Information and Technical Assistance Center. (2007). *Child Care Bulletin Issue 32*. Fairfax, VA: Child Care Bureau, Administration for Children and Families, U.S. Department of Health and Human Services.

National Child Care Information and Technical Assistance Center. (2009). *Child care and development fund report of state and territory plans FY 2008-2009*. Retrieved from http://nccic.acf.hhs.gov/pubs/stateplan2008-09/index.html

National Council on Teacher Quality. (2007). State teacher policy yearbook: Progress on teacher quality. Retrieved January 1, 2008, from http://www.nctq.org/stpy/

Norris, D. J., Dunn, L., & Eckert, L. (2003). *Reaching for the Stars: Center validation study final report*: Early Childhood Collaborative of Oklahoma.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large Are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System*. Baltimore, MD: Brookes Publishing.

Pianta, R. C., La Paro, K. M., Payne, C., Cox, M. J., & Bradley, R. (2002). The relation of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes. *The Elementary School Journal, 102*(3), 225(215).

Raudenbush, S. W., & Sampson, R. J. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology, 29*, 1-41.

Rimm-Kaufman, S. E., Curby, T. W., Grimm, K. J., Nathanson, L., & Brock, L. L. (2009). The contribution of children's self-regulation and classroom quality to children's adaptive behaviors in the kindergarten classroom. *Developmental Psychology, 45*(4), 958-972.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417-458.

Robelen, E. W. (2008, November 12). Gates' new approach gets good reviews. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2008/11/12/13gatesreact.h28.html

Robelen, E. W. (2009, January 22). Gates gives $22 million in grants. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2009/01/22/19gates.h28.html

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247-252.

Tout, K., Starr, R., & Cleveland, J. (2008). *Evaluation of Parent Aware: Minnesota's Quality Rating System pilot*. Minneapolis, MN: Minnesota Early Learning Foundation Research Consortium.

Tout, K., Zaslow, M., Halle, T., & Forry, N. (2009). *Issues for the next decade of Quality Rating and Improvement Systems* (Issue Brief No. 3). Washington, DC: Prepared by Child Trends for the Office of Planning, Research, and Evaluation, Administration for Children and Families, US Department of Health and Human Services.

U.S. Department of Health and Human Services, Administration for Children and Families, & Office of Head Start. (2008). *Classroom Assessment Scoring System* (Information Memorandum No. ACF--IM-HS-08-11). Washington, DC: Brown, Patricia E.

Zaslow, M., Tout, K., Halle, T., & Forry, N. (2009). *Multiple purposes for measuring quality in early childhood settings: Implications for collecting and communicating information on quality* (Issue Brief No. 2). Washington, DC: Prepared by Child Trends for the Office of Planning, Research, and Evaluation, Administration for Children and Families, US Department of Health and Human Services.

Zellman, G. L., & Perlman, M. (2008). *Child-care Quality Rating and Improvement Systems in five pioneer states: Implementation issues and lessons learned*. Arlington, VA: RAND Corporation.

Zellman, G. L., Perlman, M., Le, V.-N., & Setodji, C. M. (2008). *Assessing the validity of the Qualistar Early Learning Quality Rating and Improvement System as a tool for improving child-care quality*. Arlington, VA: RAND Corporation.

# Power of Two: Impact of Two Years of High Quality Teacher-Child Interactions

Anne E. Henry

University of Virginia

Kevin J. Grimm

University of California, Davis

Robert C. Pianta

University of Virginia

## Abstract

The current study uses propensity score analysis to compare the academic, social, and behavioral outcomes of children who were exposed to high quality teacher-child interactions in both pre-kindergarten and kindergarten with the outcomes of children who were exposed to low quality teacher-child interactions in both of those years. Data from the National Center for Early Development and Learning's Multi-State Study of Pre-kindergarten were analyzed. Children exposed over this two-year period to teacher-child interactions consistent with effective classroom organization and instructional support scored significantly higher on literacy and language skill assessments (effect sizes ranged from.19 to .75). Results are discussed in terms of the cumulative nature of effective teacher-child interactions across multiple years in early childhood.

## Power of Two: Impact of Two Years of High Quality Teacher-Child Interactions

Children’s experiences in pre-kindergarten and kindergarten classrooms can set a tone for academic development and school success. Trajectories for a range of outcomes become less malleable as children age and enter school, and the benefits of participation in early childhood education can be lasting (Entwisle & Alexander, 1993; Reynolds, Temple, Robertson, & Mann, 2001; Schweinhart & Weikart, 1997). High quality interactions between teachers and students in pre-kindergarten and kindergarten have been shown to predict social and academic outcomes through second grade (e.g. Burchinal et al., 2008; Peisner-Feinberg et al., 2001; Ramey et al., 2000). Features of teacher-child interactions have also been found to contribute positively to developmental outcomes for children who are likely to struggle due to behavioral and demographic risk factors (Hamre & Pianta, 2005). Still, the effects of early school experiences on achievement outcomes can fade with time (Currie & Thomas, 2000; Magnuson, Ruhm, & Waldfogel, 2007; Peisner-Feinberg et al., 2001), and many argue that continued provision of high quality educational settings through and beyond early childhood is essential for positive development (Bogard & Takanishi, 2005; Brooks-Gunn, 2003; National Association for the Education of Young Children, 2009). Using data from the National Center for Early Development and Learning’s (NCEDL) Multi-State Study of Pre-kindergarten, we examine differences in academic and social outcomes associated with experiencing high quality teacher-child interactions in pre-kindergarten and kindergarten.

A few well-known projects, including the Abecedarian Project, the Chicago Child-Parent Center Program, and the High/Scope Preschool Curriculum Comparison Study, report positive effects for at-risk children who participated in preschool interventions, particularly when the interventions incorporated additional support services for children and their families (Ramey et al., 2000; Reynolds et al., 2001; Schweinhart & Weikart, 1997). These programs were associated with better academic outcomes, higher levels of educational attainment, and lower rates of delinquent behavior in adolescence and early adulthood. Researchers and policymakers have used this information to argue for increased investment in not only pre-kindergarten programs, but in kindergarten and elementary programs that can ensure the lasting benefits of these investments (Bogard & Takanishi, 2005; Brooks-Gunn, 2003). Here, we focus on one component of early childhood education that plays an important role in children's outcomes – teacher-child interactions. We examine the effect of high quality teacher-child interactions over multiple years in early childhood. More specifically, we examine the extent to which exposure to consistently high versus consistently low quality teacher-child interactions over two years (pre-kindergarten and kindergarten) contribute to children's social and academic outcomes at the end of kindergarten.

### *Challenges to Identifying Effects*

*Measuring classroom quality and effects on outcomes.* Classroom quality can be evaluated in a variety of ways including teacher qualifications, value-added contributions to learning, or the number of books available. Yet many measures fail to capture the complexity of classroom environments. Data on teacher qualifications such as certification status or level of education have not been consistently predictive of student outcomes (e.g. Early et al., 2007).

Another way to conceptualize classroom quality is in terms of teacher-child interactions. A recent statement by the National Association for the Education of Young

Children regarding developmentally appropriate practice highlights the importance of a teacher's "moment-to-moment decisions" and interactions with children to guide learning and development (2009, p. 8). These decisions and interactions have been termed *process* quality and can be evaluated through observation (Pianta et al., 2005). Observations of teacher-child interactions have been shown to be more powerful predictors of child outcomes than teachers' perceptions of social and emotional processes, the structural/physical characteristics of classrooms, or even class size (Howes et al., 2008), and thus provide an important perspective when assessing quality.

One measure used to evaluate teacher-child interactions in early education settings is the Classroom Assessment Scoring System (CLASS; Hamre, LoCasale-Crouch, & Pianta, 2007; Hamre & Pianta, 2005; LoCasale-Crouch et al., 2007; NICHD Early Child Care Research Network, 2002, 2005; Pianta, La Paro, & Hamre, 2008b; Pianta et al., 2002). Observers using the CLASS assign global ratings that factor into three domains of interactions: Classroom Organization, Emotional Support, and Instructional Support. The 3-factor structure is based on research from education and developmental psychology, tested in more than 4,000 classrooms, and is considered generalizable to teacher-child interactions in pre-kindergarten through fifth grade (Hamre, Pianta, Mashburn, & Downer, 2008).

For children who attend pre-kindergarten and kindergarten, social, behavioral, and academic outcomes are predicted in part by the quality of the teacher-child interactions experienced in these settings (Hamre & Pianta, 2005; Howes et al., 2008; Mashburn et al., 2008; Peisner-Feinberg et al., 2001; Pianta et al., 2002). When children are exposed to high quality teacher-child interactions for a year, the effects on developmental outcomes are evident that year and some effects persist for additional years (Burchinal et al., 2008). For example, the quality of classroom practices and teacher-child relationships in pre-kindergarten predict language and math outcomes at the end of pre-kindergarten and over time (Burchinal et al., 2008; Howes et al., 2008; Mashburn et al., 2008; Peisner-Feinberg et al., 2001). Mashburn et al. (2008) found that the quality of instructional interactions in pre-kindergarten was positively associated with children's spring scores on five different measures of academic and language development. For comparison purposes with the present study, we calculated standardized regression weights for these statistically significant outcomes. Standardized betas were .04 for the Peabody Picture Vocabulary Test, .04 for the Applied Problems subtest of the Woodcock Johnson III Test of Achievement, and .06 for the Oral and Written Language Scale.[1]

The quality of emotional and instructional teacher-child interactions also predicted students' gains from fall to spring of pre-kindergarten on measures of language and literacy (Howes et al., 2008). Effect sizes (*d*) for significant outcomes ranged from .11 to .20. Comparable effects are seen for process quality in kindergarten; instructional and emotional support in kindergarten are related to competence in both language and math for the same year (Pianta et al., 2002). Some of these effects appear to persist; for example, the effect of the quality of teacher-child relationships in pre-kindergarten remains predictive of language and math outcomes through the second grade (Hamre & Pianta, 2001; Peisner-Feinberg et al., 2001). Additionally, high instructional support in pre-kindergarten is positively related to language and reading outcomes at the end of kindergarten (Burchinal et al., 2008).

Although the effect sizes mentioned above describe the impact of exposure to high quality teacher-child interactions over the pre-kindergarten *or* kindergarten year, additional work is needed to assess the impact of high quality teacher-child interactions over *multiple* years in early childhood. If language and literacy gains have effect sizes ranging from .11 to .20 given *one* year of high quality interactions in pre-kindergarten (Howes et al., 2008), effect sizes may be larger when children are exposed to two years of high quality interactions (in pre-kindergarten and kindergarten). The current paper addresses this question.

In addition to effects on achievement, instructional and emotional support in pre-kindergarten are associated with positive social and behavioral outcomes as well. Peisner-Feinberg et al. (2001) found that effective classroom practices and teacher-rated closeness with a child in pre-kindergarten predicted increases in child sociability through kindergarten and decreased problem behaviors through the second grade. The quality of emotional interactions between teachers and students during pre-kindergarten is positively associated with teachers' reports of students' social competence at the end of the year, and negatively associated with reports of problem behaviors (Mashburn et al., 2008; NICHD Early Child Care Research Network, 2002). Kindergarten classrooms characterized by high level instruction, child-centered activities, and an emotionally supportive teacher are associated with kindergarteners' observed on-task behavior (Pianta et al., 2002). Additionally, the quality of teacher-child relationships in kindergarten predicts social skills through second grade (Peisner-Feinberg et al., 2001). Again we wonder, however, if the effects of high quality interactions on social and behavioral outcomes would be even greater if children were exposed to multiple years of high quality support.

Teacher-child interactions that foster an organized classroom are important for creating an environment where students become engaged in the learning process (Emmer & Stough, 2001) and are associated with students' academic, social, and behavioral outcomes. In one study, teachers whose students had high levels of engagement, reading ability, and writing skills were observed to frequently monitor their classrooms, encourage students to problem-solve and stay on task, and conduct well-planned lessons (Pressley et al., 2001). In early childhood, the presence of chaos in the classroom (and thus the absence of control or organization) is negatively associated with children's observed compliance (Wachs, Gurkas, & Kontos, 2004). Classroom organization has been shown to have a greater impact on child outcomes when it is proactive as opposed to reactive. Students with teachers who spent more time organizing classroom rules and routines in the fall and less time later in the year showed significantly greater gains in word reading skills than their peers with teachers who spent less time organizing in the fall and the same or increased amounts of time thereafter (Cameron, Connor, Morrison, & Jewkes, 2008). However, questions remain regarding the impact on children's outcomes when they are exposed to multiple years of teachers who effectively engage them and manage their time.

As we have mentioned, estimates of the impact of teacher-child interactions on child outcomes are limited by the reality of children's experiences across multiple years. Unfortunately, individual children rarely experience consistently high quality interactions with teachers from one year to the next in early childhood (La Paro et al., in press;

NICHD Early Child Care Research Network, 2005; Peisner-Feinberg et al., 2001; Pianta, Belsky, Houts, Morrison, & NICHD ECCRN, 2007). This is not surprising given the large variation in quality among teachers, even teachers within the same school (Nye et al., 2004; Rivkin et al., 2005; Rockoff, 2004). Still, we know that continuity in high quality interactions across years is important for children's development. For example, in the National Institute of Child Health and Human Development (NICHD) Study of Early Child Care, increased achievement at 54 months was associated with observed quality in child care between 6-24 months and 36-54 months, and the gains were greater if observed quality was high over this entire period of time (NICHD Early Child Care Research Network & Duncan, 2003). We examine whether academic and social outcomes are greater for children when observed interactions are high quality over multiple years in pre-kindergarten and kindergarten.

*Selection bias.* Adding to the challenges of estimating the causal impact of teacher-child interactions on children's development is the fact that students who do experience high quality teacher-child interactions across multiple years in early childhood can have very different characteristics than students who do not. The term *selection bias* is used when treatment assignment leads to a correlation between assignment and outcomes in the absence of treatment (B. D. Meyer, 1995). In this case, we consider the experience of high quality interactions to be the "treatment," and bias is present because student characteristics are associated with likelihood of exposure to high quality interactions. Selection bias is a major barrier to inferring causality in studies of classroom effects because teachers are not randomly assigned to schools and students are rarely randomly assigned to schools or teachers. The result is that advantaged students tend to end up in classrooms with higher qualified teachers and students at risk for poor achievement tend to end up in classrooms with less qualified teachers (Clotfelter, Ladd, Vigdor, & Wheeler, 2006; Clotfelter, Ladd, & Vigdor, 2006; Lankford, Loeb, & Wyckoff, 2002). Certain demographic characteristics relate to children's likelihood of experiencing high or low classroom quality, such as family income and race (Hamre & Pianta, 2005; LoCasale-Crouch et al., 2007; Magnuson & Waldfogel, 2005; Peisner-Feinberg et al., 2001; Pianta et al., 2005; Pianta et al., 2002). When substantial differences exist between children who are exposed to high quality classrooms and children exposed to low quality classrooms, distinguishing the effect of high quality interactions above and beyond these differences is difficult.

In the current study, we estimate the impact of two years of high quality teacher-child interactions in pre-kindergarten and kindergarten programs that are operating at-scale. Given the ethical constraints of randomly assigning children to teachers who interact with them in high or low quality ways, we must work within the framework of the interactions that have already been observed in classrooms. In turn, we must attempt to statistically eliminate selection bias.

### *Addressing Selection Bias*

There are a number of ways to account for selection bias when estimating treatment effects. When random assignment is not possible, the simplest and most common technique to account for selection bias is the use of covariates in a regression framework. Outcomes are regressed on treatment condition, controlling for covariates which may include individual characteristics or environmental factors. With this technique, however, there is no way to be sure that unobserved covariates are not also introducing selection bias (Burchinal et al., 2008). In assessing the relationship between pre-kindergarten quality and kindergarten outcomes, Burchinal et al. acknowledge that the covariates used in the reported regression analyses were unlikely to completely remove selection bias due to the correlation between pre-kindergarten quality and children's skills at pre-kindergarten entry. In this paper, we attempt to further reduce selection bias using an alternative approach, propensity score matching.

Alternative methods for reducing selection bias have been developed in recent decades (Shadish, Cook, & Campbell, 2002). For example, nonequivalent comparison groups can be used to calculate estimates comparable to those achieved with randomly formed control groups (Aiken, West, Schwalm, Carroll, & Hsiung, 1998; Michalopoulos, Bloom, & Hill, 2004). In this paper, maximally similar nonequivalent comparison groups are created by matching on propensity scores that represent individuals' probability of assignment to treatment based on a vector of observed characteristics.

Rosenbaum and Rubin (1983) introduced propensity scores, arguing that the scores work just as well for matching comparison groups as matching on all covariates from which the scores are calculated. Recent comparisons of experimental and propensity score designs have demonstrated that the methods achieve similar effects (Dehejia & Wahba, 1999; Shadish, Clark, & Steiner, 2008). For example, Shadish et al. (2008) randomized participants to be in either a randomized or nonrandomized experiment. Propensity score methods reduced bias in the nonrandomized study by 58 to 96% when covariate-adjusted randomized results were used as the reference. The percentage of bias reduced depended upon the outcome and adjustment method.

There are a few situations when propensity score techniques are preferable to ordinary linear regression. For one, when using propensity score matching or stratification, selection bias can be reduced by avoiding some of the linearity assumptions of ordinary linear regression (Shadish et al., 2008; J. Smith, 2000). Second, propensity score matching is preferable when designing treatment and control groups that are aligned on a large number of covariates (Shadish et al., 2008; J. Smith, 2000). For some values of observed covariates there will not be close matches between treatment and comparison group observations; this is less of a problem when matching is done with propensity scores, which are scalar. In the current study, we wanted to design treatment and comparison groups from secondary data which was very rich in possible covariates. For this reason, we selected propensity score matching as the analytic technique.

There are numerous matching schemes available, and most involve nearest-neighbor matching, where the treatment and comparison observations are paired based on who has the closest score. Comparisons of estimates from nearest-neighbor methods indicate that bias is effectively reduced when matching within propensity score calipers followed by Mahalanobis metric matching on key covariates (Rosenbaum & Rubin, 1985; Rubin & Thomas, 2000).

When comparison groups are created using matching or stratification on propensity scores, the comparability of the groups can be tested. First, it is important to establish that there are propensity score values for which there are observations in both samples (J. Smith, 2000). Second, it is important for researchers to assess the comparability of treatment and comparison groups on all observed covariates. If the distribution of covariates is balanced within each group, observed differences in outcomes can be attributed to the effect of treatment and not differences in observed covariates (Rubin, 2001). This assumption is already present in randomized experiments. When randomization takes place, treatment and comparison participants have the same, fully known, propensity scores and the distribution of covariates is assumed to be equal for treatment and comparison groups, for both observed and unobserved covariates. Differences between groups in true randomized experiments are interpreted as effects of treatment. With propensity score matching, the goal is to create groups for which the distribution of observed covariates is maximally similar. Although matching on unobserved covariates is impossible, propensity score matching is informed by all observed covariates that are theoretically relevant. Rubin (2001) outlined three conditions for assessing whether matching has reduced selection bias and created groups with balanced covariates. These guidelines concern the means and variances of the propensity scores and the covariates from which they are constructed.

### *The Current Study*

In the present study we use propensity scores to create groups of children who were equally likely to be exposed to high quality teacher-child interactions yet differed in their actual experience of supportive interactions in an effort to isolate the effect of this specific aspect of early education programs. We test whether the high quality and low quality groups we create have similar distributions of propensity scores and observed covariates following Rubin’s (2001) recommendations for comparability. Once comparability of groups is established, differences at the end of kindergarten can be interpreted as the impact of quality experienced over two years. We specifically compare the academic, social, and behavioral outcomes of children who were exposed to high quality teacher-child interactions in both pre-kindergarten and kindergarten with the outcomes of children who were exposed to low quality interactions for these two years. These differences describe the potential effect of exposing children to multiple years of high quality teacher-child interactions, irrespective of their background characteristics at the beginning of pre-kindergarten.

Classroom quality was measured through observations of teacher-child interactions in three domains: Classroom Organization, Instructional Support, and Emotional Support. Because children exposed to high quality interactions could have different characteristics than children exposed to low quality interactions, propensity score matching was used to create groups of students who, at the beginning of pre-kindergarten, observably differed only in their experience of quality. At the end of kindergarten, children who experienced two years of high quality organizational support are expected to present fewer behavior problems, children exposed to quality emotional interactions are expected to be more socially competent, and children exposed to high quality instruction are expected to have higher scores on tests of academic achievement.

## Method

### *Sample*

The data come from the National Center for Early Development and Learning's (NCEDL) Multi-State Study of Pre-kindergarten. All pre-kindergarten classrooms in this study were center-based programs with full or partial state funding and direction. Prior to data collection in 2001-2003, selection of classrooms and children began at the state level. Investigators selected six states that represented variability in length of school day, teacher credentialing requirements, school locations, and geography. These states were selected from a larger pool of potential states that served at least 15,000 (15%) 4-year-olds.

Following state selection, researchers randomly selected 20 zip codes within each state. Next, two sites receiving pre-kindergarten funding were randomly selected within each zip code. Finally, researchers randomly selected one classroom and four children within that classroom for each site. Researchers collected data from a total of 240 classrooms and 960 children in pre-kindergarten. Complete data were collected for 778 children in over 800 classrooms for kindergarten data collection, and partial data were available for another 132 kindergarteners. We excluded children given Spanish assessments at any point from analyses in the current study. The final sample for this study included 777 children who were never assessed in Spanish, and whose classrooms were observed in both pre-kindergarten and kindergarten.

About half of the students were male (49%) and the sample was ethnically diverse with 45% White, 27% African-American, 15% Hispanic, 8% Multi-racial, 3% Asian/Pacific-Islander, 1% Native American. Maternal education varied with the largest proportion reporting high school (46%) as their highest education level, followed by 14% who did not finish high school. Just over half (53%) of the children belonged to families whose annual incomes were less than or equal to 150% of the federal poverty income guidelines for their family's size.

### *Measures*

*Quality of teacher-child interactions.* The Classroom Assessment Scoring System (CLASS; Pianta, La Paro et al., 2008b) was used to make observations of teacher-child interactions in both pre-kindergarten and kindergarten. Researchers have validated the

CLASS through standardized observations in four large-scale projects involving over 4,000 classrooms (Hamre et al., 2008). Research has shown that the global ratings from this scale are generalizable to teacher-child interactions from pre-kindergarten through the fifth grade and predict to child outcomes across these grades (Burchinal et al., 2008; Mashburn et al., 2008).

In the present study, the CLASS assessed nine dimensions of teacher-child interactions, each assigned a score from 1-7. Scores of 1-2 are considered low quality, 3-5 are mid-range, and 6-7 are considered high quality. The dimensions are divided into three domains. *Classroom Organization* (comprised of dimensions Behavior Management, Productivity, and Instructional Learning Formats) describes clear behavioral expectations, established routines to maximize learning opportunities, and teacher-facilitated exploration of learning materials. *Emotional Support* (dimensions include Positive Climate; Teacher Sensitivity; Over-Control, reverse-scored), includes children's contact with positive, sensitive, and responsive teachers. *Instructional Support* (dimensions are Concept Development and Quality of Feedback) is characterized by focused discussion, purposeful scaffolding, and specific feedback.

Data collectors were tested for reliability in the fall and spring of each year, and were considered reliable if 80% of their scores were within one point of the gold standard response. For four reliability tests, data collectors' reliability ranged from 86% to 93%. The mean weighted kappa ranged from .60 to .73.

Data collectors made CLASS observations over two days in the fall and two days in the spring in the pre-kindergarten year. In kindergarten, CLASS observations occurred three times across the school year, with at least four weeks between observations. Observers rated classrooms on all dimensions every 30 minutes throughout the day.

*Child outcomes.* The Peabody Picture Vocabulary Test, 3rd edition (PPVT; Dunn & Dunn, 1997) was administered in the fall and spring of both years. In this test of receptive vocabulary, children selected which of 4 pictures matched the word spoken by the examiner. We used standard scores for the current paper. The median reliability of the items on this scale is .94.

The Woodcock-Johnson III Tests of Achievement (Woodcock, McGrew, & Mather, 2001) was administered in the fall and spring of each year. Academic achievement was assessed with three subtests from this battery. Applied Problems was used to assess skill in solving math problems; reliability coefficients for this subtest ranged from .92 to .94 for 3- to 5-year-olds. The Sound Awareness / Rhyming subtest was used to examine rhyming ability; reliability coefficients ranged from .71 to .85 for 4- to 5-year-olds. Letter-Word Identification was used to measure children's ability to identify words and letters; reliability coefficients ranged from .97 to .99 for 3- to 5-year-olds.

The Oral and Written Language Scale (OWLS; Carrow-Woolfolk, 1995) was administered in the fall and spring of each year to assess understanding and use of spoken language. Test-retest reliability for the 4-to 5-year-old age range was .86.

In letter naming, children viewed a set of mixed capital and lowercase letters and identified as many as possible. In number naming, children viewed a sheet of numbers from 1 to 10, printed in random order, and identified as many as possible. Children were also asked to count and point, with one-to-one correspondence, to a picture of 20 teddy bears. Children who counted correctly were presented with a second picture for a maximum score of 40.

Teachers rated children's skills on the Academic Rating Scale (National Center for Education Statistics, 1994). They rated language and literacy items in the fall and spring of both years on a scale from 1 (*not yet*) to 5 (*proficient*), and the reliability of this measure was high (α = .89 - .94 over this time period). Mathematical reasoning items were rated in the spring of pre-kindergarten, and the fall and spring of kindergarten (α = .94).

Teachers rated students' social skills and behavior problems in the fall and spring of each year using the Teacher Child Rating Scale (Hightower et al., 1986). Items reflecting children's social competence described their assertiveness, peer social skills, task orientation, and frustration tolerance. The scale for these items was from 1 (*not at all*) to 5 (*very well*). The reliability coefficients for social skills ranged from .94 to .96 over the course of data collection. Items reflecting behavior problems described children's conduct, internalizing, and learning problems. The response scale for these items was from 1 (*not at problem*) to 5 (*very serious problem*). The internal consistency among the behavior problem items was high (α=.92 from fall of pre-kindergarten through spring of kindergarten).

Teachers also rated perceptions of their relationship with each child in the spring of both years, using the Student Teacher Relationship Scale (STRS; Pianta, 2001). The STRS is a valid predictor of academic and social functioning in pre-kindergarten through the elementary grades (Pianta et al., 2002). Items were rated from 1 (*definitely does not* *apply*) to 5 (*definitely applies*), and generated two relationship quality scores, Closeness ($\alpha$=.85) and Conflict ($\alpha$=.87).

*Data Analysis*

Analyses took place in four steps: (1) Imputation was conducted to address missing data. (2) Students who were exposed to consistently high or consistently low quality teacher-child interactions in pre-kindergarten and kindergarten were grouped accordingly. (3) To address selection bias in estimating effects, propensity score matching was used to create two groups of students who had an equal likelihood of experiencing high quality teacher-child interactions in pre-kindergarten and kindergarten. (4) Regression analyses were conducted to determine the extent to which group membership was associated with child outcomes in the spring of kindergarten. See Appendix A for a flow chart illustrating the analytic process.

Of 777 children whose classrooms were observed in pre-kindergarten and kindergarten, analyses focused on children who experienced high quality interactions with teachers in both years matched with children who experienced low quality teacher-child interactions. Although 777 children were included, a limited number of these children experienced consistent quality over this time period. This was not surprising given that individual children rarely experience consistent educational supports across years (La Paro et al., in press).

Likewise, children who experienced a consistent quality level on one CLASS domain did not necessarily experience the same level of quality on other CLASS domains. *Therefore, analyses were run separately for each of the CLASS domains of Classroom Organization, Emotional Support, and Instructional Support, which resulted* *in a different subsample of children for each.* The above steps 2-4 were repeated for groups based on each of the CLASS domains. The results were interpreted separately for each CLASS domain; because a different subsample was used for each, caution is warranted in comparing results across domains.

*Group assignment.* First, classrooms were identified as presenting low (1 or 2), mid (3-5), or high (6 or 7) quality teacher-child interactions. One CLASS domain score per classroom was calculated by averaging CLASS ratings across all observations for a given year. Because few observed classrooms were rated in the upper end of the CLASS scale (6 or 7 out of 7 possible points), low and high quality classrooms were identified by their domain scores relative to the distribution for observed classrooms in the study, using terciles. We categorized classrooms scoring in the bottom third as low quality, and classrooms in the top third as high quality. Table 1 details the domain cutoff scores used to assign classrooms as low or high quality. Importantly, the scores are dependent on the level of quality that existed in the pre-kindergarten and kindergarten classrooms observed. This approach was considered conservative because ratings of existing teacher-child interactions averaged in the mid- or low-range in some cases. For example, “high” Instructional Support was defined here as an average score above 2.42 in pre-kindergarten and 2.06 in kindergarten. These scores are actually considered to be low quality on the standard CLASS scale. Given effects on child outcomes using this definition, we would expect even greater effects for children who experience high quality (i.e., a score of 6 or 7) based on the CLASS scale. Also, the scores for low and high are not very different from each other in any of the three domains. A score of 4.34 was considered low quality Classroom Organization in kindergarten, when a score of 4.99

was considered high. Any differences between the low and high quality groups were contingent on this very fine distinction.

Once classrooms representing low or high quality interactions were selected, students who were in high or low quality classrooms over both years were identified. Consistently high quality was considered exposure to “treatment” and consistently low quality the “comparison”. It would be inappropriate to compare these two groups without making adjustments for selection as these groups significantly differed in terms of child characteristics, achievement, and behavior. Propensity scores were calculated to create two matched groups to control for selection effects.

*Propensity scores.* Propensity scores were calculated via logistic regression. Covariates were selected that were both theoretically important to the selection process and correlated with outcomes. In this study, data gathered in the fall of pre-kindergarten were entered into a logistic regression predicting membership in the treatment (consistently high quality) versus the comparison (consistently low quality) group. Covariates were included as main effects and reflected child’s gender, age, family income, ethnicity, state, child care environment prior to pre-kindergarten, the number of people living in the home, and whether or not the child’s father lived at home. Propensity scores were children’s predicted probabilities of group membership based on this model. Once created, propensity scores were transformed to the logit scale (Rubin, 2001).

*Matching on propensity scores.* Propensity scores were used to match students who consistently experienced high quality interactions with students who consistently experienced low quality interactions. A SAS Macro was used to match students in the high quality group to students in the low quality group (Feng, Jun, & Xu, 2006). One case at a time, a case from the low quality group (comparison group) was selected and its propensity score was compared to the propensity scores of cases in the high quality (treatment) group. When there was one case in the treatment group for which the propensity score was within the caliper (one quarter of the standard deviation of the logit), then this case was selected from the treatment group as a match for the comparison. When more than one possible match from the treatment group was available (i.e. at least two treatment cases with propensity scores within the caliper), then the closest match was identified according to each case's Mahalanobis distance from the comparison case. When there was not a match in the treatment group for which the propensity score was within the caliper of the comparison group propensity score, then the comparison case was dropped from the following analyses. Appendix B provides a visual representation of the matching process.

*Assessing bias reduction and balance.* Once the treatment and comparison groups had been matched, the degree to which bias was reduced was examined. Rubin (2001) outlines three conditions that must be met to establish comparability of groups: (1) The differences in the means of the propensity scores in the treatment and comparison groups must be less than half a standard deviation apart. (2) The ratio of the variances of the propensity score in the treatment and comparison group must be close to one. (3) Each covariate must be regressed on the logit of the propensity score and the ratio of the variances of the residuals from each of these regressions must be close to one. If any of the three conditions were not met, covariates were added to and sometimes removed from the initial logistic regression until balance improved. As a final check, pretest differences between treatment and comparison groups were examined.

*Estimating effects.* Between-group mean differences were assessed to determine the effect of treatment on academic and social outcomes in the spring of kindergarten. Effect sizes were calculated to facilitate comparison across domains of interaction. Cohen's *d*, the difference in means divided by the pooled standard deviation, was selected for this purpose. Pooled standard deviations were calculated for each of the outcomes from the full sample of 777 students prior to group assignment and matching procedures because the matched pairs for each domain reflected a different subsample of the NCEDL data.

## Results

### *Group Assignment*

Using the CLASS cutoff scores detailed in Table 1, children were identified who were exposed to two years of high quality or two years of low quality teacher-child interactions. High (treatment) and low (comparison) quality groups were established for each of the three CLASS domains – Classroom Organization, Emotional Support, and Instructional Support. As expected, a limited number of children experienced consistently high or consistently low quality. The subsample sizes for each of the three CLASS domains were smaller than the original sample of 777 children (*n*=192, 220, and 207 for Organization, Emotional, and Instructional respectively).

### *Creating Propensity Scores*

Propensity scores were calculated in three logistic regressions, one for each of the three CLASS domains. Ultimately, the same 17 covariates were used in each of these regressions and reflected characteristics of the child (gender, age in the fall of prekindergarten, whether or not the child was Black or Hispanic), characteristics of the family (income less than or equal to 150% of the federal poverty income guidelines, number of people in the household, whether or not the father lives in the household), child care arrangements in previous years (was child in Head Start, a child care center, in the same prekindergarten as the study year, or at home), and dummy codes representing the 6 states where data was collected. There were several steps to assess comparability of groups prior to settling on this final list of covariates, including examination of propensity score distributions, means, and variances.

### *Assessing Bias Reduction and Balance*

Histograms illustrating the distribution of propensity scores pre- and post-match for each of the CLASS domains were examined (see Appendix C). Pre-match histograms verify that selection bias was present. Children exposed to high quality teacher-child interactions over two years had higher propensity scores meaning they were more likely to belong to the treatment groups. Still, there were propensity score values for which there were observations in both treatment and comparison groups, further illustrated in post-match histograms. A sufficient number of children were equally likely to experience high quality teacher-child interactions, regardless of whether they actually experienced high quality interactions.

As described previously, Rubin (2001) noted three additional guidelines for testing group comparability. These conditions were not met prior to matching but they were met for the final matched samples. For all three CLASS domains, the degree of balance and bias reduction was sufficient to establish comparability after matching (see Table 2).

In a final comparability test, groups were not different on standardized tests or teachers' reports of academic and social skills in the fall of pre-kindergarten. Significant differences between groups were found in just two instances. Children in the low and high quality Classroom Organization groups had significantly different scores on the Rhyming subtest of the Woodcock Johnson. Also, teacher-reported language skills were significantly different between the low and high quality Emotional Support groups in the fall of pre-kindergarten. Further analyses for outcomes at the end of kindergarten were not conducted for these two subtests.

### *Matched Sample Characteristics*

The final matched samples were much smaller than the original sample for two reasons. First, only some of the original 777 children were exposed to two years of high or two years of low quality interactions on any of the CLASS domains. Second, only some of those children had propensity scores similar enough to be matched across groups. The final sample for Classroom Organization included 56 children (28 matched pairs), Emotional Support included 106 children (53 pairs), and Instructional Support included 90 children (45 pairs).

The matched samples were very similar to the full sample in terms of gender, race/ethnicity, maternal education, and level of poverty. In some cases, the same children were represented in the treatment or control groups for more than one domain. For the most part, less than 20% of the sample for each domain was represented in one of the other domains. There were two exceptions: 45% of the children in the Classroom Organization sample were also members of the Emotional Support sample, and 25% of the children in the Emotional Support sample were also members of the Classroom Organization sample.

*Estimating Effects*

Given the equality of the treatment and comparison groups in the fall of pre-kindergarten, differences at the end of kindergarten were interpreted as the impact of each of three domains of the quality of teacher-child interactions experienced over the two years. Unstandardized regression coefficients, significance, standard errors, and effect sizes (*d*) for each CLASS domain are presented in Table 3. Regression coefficients represent the mean difference between scores for children exposed to two years of low quality interactions in each of three domains and children exposed to two years of high quality interactions in each of those three domains.

*Classroom Organization.* At the end of kindergarten, children who experienced high levels of Classroom Organization in both years performed significantly better than their peers experiencing low levels of organization on several direct assessments of their skills. The average score for children in the treatment group on the Peabody Picture Vocabulary Test (PPVT) was almost 9 points higher than children in the comparison group; this was statistically significant (*p*=.01) and had a large effect size of .75. Children in the treatment group named significantly more capital and lowercase letters on a test of naming letters (*p* = .03, *d=.59*). There was also a trend toward significance for these children to score higher on the Letter-Word Identification subtest of the Woodcock Johnson.

Teachers rated children exposed to high levels of Classroom Organization as having significantly higher language skills and there was a trend for higher teacher ratings of math skills. Several outcomes were not statistically significant, yet had effect sizes greater than .30 and are worth mentioning. Children exposed to high quality organizational support could identify more numbers, but were able to count with one-to-one correspondence for fewer of them. Children in classrooms with high levels of Classroom Organization were rated as having fewer behavior problems and greater social competence.

*Emotional Support.* In classrooms providing high levels of Emotional Support, there was a trend that teachers reported fewer perceptions of conflict with their students. Though not statistically significant, students who experienced high levels of emotional support in both years had better language and literacy skills at the end of kindergarten, as indicated by higher scores on the PPVT. The effect size for the PPVT was moderate at .33.

*Instructional Support.* In the spring of kindergarten, students who experienced high levels of Instructional Support over two years had significantly higher performance on the Rhyming subtest of the Woodcock Johnson, with an effect size of .45. There was a trend toward significance for the Oral and Written Language Scale. Of mention, but not significant, was the difference in scores on the PPVT. Children in classrooms providing high levels of instructional support over two years had an average score 4 points higher than children in low quality classrooms, representing an effect size of .35. There were no significant differences in teacher report of student outcomes for this domain of interactions.

## Discussion

In this paper, we estimated the impact of exposure to two years of high quality teacher-child interactions in pre-kindergarten and kindergarten on child outcomes at the end of kindergarten. We reduced selection bias by constructing groups of children who were similar in their likelihood of exposure to high quality teacher-child interactions, and could thus infer that differences in outcomes at the end of kindergarten were the effect of the Classroom Organization, Emotional Support, or Instructional Support that children were exposed to. Because this is an observational study instead of a randomized experiment, we were limited by the range of teacher-child interaction quality present in this sample. Still, statistically significant differences in outcomes had effect sizes that ranged from .32 to .75 in terms of benefits to child outcomes.

Supportive organizational and instructional interactions over two years predicted significantly higher performance on several measures of children's language and literacy skills. Some of these effects were quite large. For example, high quality Classroom Organization led to a 9-point gain on the PPVT - more than half of a standard deviation. The results for Instructional Support were consistent with findings from Burchinal et al. (2008), who found that Instructional Support in pre-kindergarten significantly predicted language outcomes but not math outcomes. Across all of the results, greater exposure to high quality teacher-child interactions predicted scores in the expected directions. Larger effect sizes and significance occurred for outcomes evaluated via standardized assessments, indicating that the interactions which occurred in these classrooms were impacting children's actual skills, not just teachers' perceptions of these skills.

The findings are most striking when considering that these effects directly resulted from the experience of quality teacher-child interactions. For each domain, those children who experienced high or low quality were equal in terms of selection characteristics and baseline performance. Matched samples met Rubin's guidelines for closeness (2001). Because the distribution of observed covariates is balanced within each group, selection bias is reduced and observed differences in outcomes are not due to differences in these covariates (Rubin, 2001). Assuming that all relevant covariates have been observed, differences in outcomes represent the effect of treatment. In this case, we can interpret differences at the end of kindergarten as an estimate of the causal effect of teacher-child interactions experienced over the two years.

The effects in this paper describe possible outcomes for children when they are exposed to consistently high quality teacher-child interactions across multiple years in early childhood. One way to interpret these results is in the language of the achievement gap. Rock and Stenner (2005) reviewed estimates of gaps in school readiness by race. The estimates they reported for the black-white gap on the PPVT ranged from 1.14 – 1.71 unadjusted standard deviations, and .69 - .95 standard deviations when adjusted for income, head of household, maternal age and education, and home environment. Here, we see that high quality Classroom Organization over pre-kindergarten and kindergarten leads to .75 standard deviation gain in PPVT scores – almost closing the gap when the adjusted estimates are referenced. Emotional and Instructional Support are also influential, each potentially narrowing the gap by a third of a standard deviation. If the racial/ethnic gap is estimated at a more conservative .5 SD, even an effect size of .3 as we see for many of our outcomes is important. Just .3 standard deviations change narrows the gap by 2-14%, depending upon the population targeted for support (Magnuson & Waldfogel, 2005).

Unfortunately, as we have seen here, children who fall on the disadvantaged end of the achievement gap are unlikely to be exposed to high quality teacher-child interactions in both pre-kindergarten and kindergarten. In fact other reports note that exposure to consistently high quality teacher interactions is rather rare even in normative, non-risk samples (Pianta et al., 2007). The selection bias that is so difficult to remove in studies of teacher quality and the likelihood of a child moving into a rather poor or mediocre classroom in a subsequent year are realities for early childhood programs operating at-scale. The achievement gap is likely to increase during this time period unless children's likelihood of exposure to high quality teacher-child interactions is systematically improved in a given year and that a multi-year pipeline of effective classroom experience is provided.

### *Limitations*

There are a few possible reasons why more of the differences did not reach statistical significance. First, the average quality of interactions in each of the domains was not very different between treatment and comparison groups. In fact, the quality scores averaged within two points of each other for the low and high quality groups in all domains. Second, although the groups experiencing high quality teacher-child interactions represent the top third of observed quality scores over the two years, the scores associated with these groups were not considered very high quality. For example, the mean Instructional Support scores for the high instructional quality group are 3.08 in pre-kindergarten and 2.65 in kindergarten. On the 7-point scale, these scores actually represent mid-level or low quality. Given this limitation, the presence of any significant findings at all becomes more meaningful. We would expect even larger effects on children’s outcomes if interactions were observed in the high range (6 or 7 of 7 possible points).

Although the study was limited by small differences between low and high quality groups in terms of experienced interactions, sample size limited the possibility of creating groups that were more different on these interactions. The small number of pairs examined per domain speaks to the low likelihood of actually experiencing high quality over time. One obstacle to studying the effects of exposure to multiple years of high quality classrooms is that students are unlikely to experience consistent quality over time, much less consistently high quality (Currie & Thomas, 2000; La Paro et al., in press). This is certainly the case for the sample used here. Due to the small final sample size in each of our analyses, effect sizes should be interpreted with caution. Effect sizes, while not directly related to sample size, can deviate farther from the population effect size with small samples (Fan, 2001).

The small sample sizes also remind us that particular children were chosen to test effects in each domain. These samples do not include children who are very unlikely to experience high quality teacher-child interactions, or children who are very likely to receive this support. Instead, we have examined the effects of supportive interactions for children who fall in the middle of the propensity score distribution – children who are the most similar to each other in terms of their background characteristics. This study is quasi-experimental; only observed characteristics are accounted for in our estimation of effects and it is possible that some selection bias remains.

### *Conclusion and Future Directions*

Past research indicates that children who are most vulnerable to academic, social or behavior problems benefit the most from experiencing high quality interactions in the classroom (Hamre & Pianta, 2005). The current paper equalizes students in terms of vulnerability. Future work should consider how the interaction of vulnerability and quality experienced over time impacts child outcomes. Alternatively, would teacher-child interactions have a greater impact on student outcomes if high quality was sustained over a longer period of time? Or if the quality experienced was of an even higher degree?

As researchers, policymakers, and administrators continue to invest in early childhood programming, they must consider the outcomes resulting from quality teacher-child interactions in these settings. In the current paper, we see academic benefits for students who are consistently exposed to pre-kindergarten and kindergarten classrooms characterized by high levels of organizational, emotional, and instructional supports. High quality teacher-child interactions over time play an important role in the development of all students, regardless of their characteristics at school entry.

# References

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J. L., & Hsiung, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review, 22*(2), 207.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, D.C.: AERA.

Arnett, J. (1989). Caregivers in day-care centers: Does training matter. *Journal of Applied Developmental Psychology, 10*(4), 541-552.

Barnett, W. S., Epstein, D. J., Friedman, A. H., Boyd, J. S., & Hustedt, J. T. (2008). *The state of preschool 2008: State preschool yearbook*. New Brunswick, NJ: National Institute for Early Education Research, Rutgers University.

Bogard, K., & Takanishi, R. (2005). PK-3: An aligned and coordinated approach to education for children 3 to 8 years old. *Social Policy Report, 19*(3), 1-21.

Boyd, D., Goldhaber, D., Lankford, H., & Wyckoff, J. (2007). The Effect of Certification and Preparation on Teacher Quality. *Future of Children, 17*(1), 45-68.

Bradley, R. H., Caldwell, B. M., & Corwyn, R. F. (2003). The Child Care HOME Inventories: assessing the quality of family child care homes. *Early Childhood Research Quarterly, 18*(3), 294-309.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brooks-Gunn, J. (2003). Do you believe in magic? What we can expect from early childhood intervention programs. *Social Policy Report, 17*(1), 1-14.

Bryant, D. M. (2000). *Validating North Carolina's 5-star child care licensing system*. Chapel Hill, NC: Frank Porter Graham Child Development Center.

Bryant, D. M., Clifford, R. M., & Peisner, E. S. (1991). Best Practices for Beginners: Developmental Appropriateness in Kindergarten. *American Educational Research Journal, 28*(4), 783.

Burchinal, M., Howes, C., Pianta, R. C., Bryant, D., Early, D. M., Clifford, R. M., et al. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher–child interactions and instruction. *Applied Developmental Science, 12*(3), 140-153.

Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2009). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly*.

Cameron, C. E., Connor, C. M. D., Morrison, F. J., & Jewkes, A. M. (2008). Effects of classroom organization on letter–word reading in first grade. *Journal of School Psychology, 46*(2), 173-192.

Carrow-Woolfolk, E. (1995). *Oral and Written Language Scales (OWLS)*. Circle Pines, MN: American Guidance Service.

Clifford, R. M., Barbarin, O., Chang, F., Early, D., Bryant, D., Howes, C., et al. (2005). What is Pre-Kindergarten? Characteristics of Public Pre-Kindergarten Programs. *Applied Developmental Science, 9*(3), 126-143.

Clotfelter, C. T., Ladd, H. F., Vigdor, J., & Wheeler, J. (2006). High-poverty schools and the distribution of teachers and principals. *North Carolina Law Review, 85*, 1345.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources, 41*(4), 778.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review, 26*(6), 673-682.

Colvin, G., Flannery, K. B., Sugai, G., & Monegan, J. (2009). Using Observational Data to Provide Performance Feedback to Teachers: A High School Case Study. *Preventing School Failure, 53*(2), 95-104.

Connor, C. M. D., Morrison, F. J., Fishman, B. J., Ponitz, C. C., Glasney, S., Underwood, P. S., et al. (2009). The ISI Classroom Observation System: Examining the Literacy Instruction Provided to Individual Students. *Educational Researcher, 38*(2), 85.

Connor, C. M. D., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., et al. (2009). Individualizing student instruction precisely: Effects of child by instruction interactions on first graders' literacy development. *Child Development, 80*(1), 77–100.

Connor, C. M. D., Son, S. H., Hindman, A. H., & Morrison, F. J. (2005). Teacher qualifications, classroom practices, family characteristics, and preschool experience: Complex effects on first graders' vocabulary and early reading outcomes. *Journal of School Psychology, 43*(4), 343-375.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Curby, T. W., Brock, L. L., & Hamre, B. K. (2009). The role of consistency in preschool teacher-child interactions. *Manuscript under review*.

Curby, T. W., LoCasale-Crouch, J., Konold, T. R., Pianta, R. C., Howes, C., Burchinal, M., et al. (2009). The Relations of Observed Pre-K Classroom Quality Profiles to Children's Achievement and Social Competence. *Early Education & Development*, *20*(2), 346-372.

Currie, J., & Thomas, D. (2000). School quality and the longer-term effects of Head Start. *The Journal of Human Resources*, *35*(4), 755-774.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Davidson, M. R., Fields, M. K., & Yang, J. (2009). A randomized trial study of a preschool literacy curriculum: The importance of implementation. *Journal of Research on Educational Effectiveness*, *2*(3), 177-208.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association, 94*(448), 1053-1062.

Dickinson, D. K., & Caswell, L. (2007). Building support for language and early literacy in preschool classrooms through in-service professional development: Effects of the Literacy Environment Enrichment Program (LEEP). *Early Childhood Research Quarterly, 22*(2), 243-260.

Domitrovich, C. E., Greenberg, m. T., Kusche, C., & Cortes, R. (2004). *The preschool PATHS curriculum*. State College, PA: Pennsylvania State University.

Downer, J. T., & Hamre, B. K. (2010). *Beliefs about intentional instruction*: Unpublished measure.

Downey, C. J., English, F. W., Steffy, B. E., Poston Jr, W. K., & Frase, L. E. (2004). *The three-minute classroom walk-through: Changing school supervisory practice one teacher at a time*. Thousand Oaks, CA: Corwin Press.

Duncan, A. (2009). *The early learning challenge: Raising the bar*. Retrieved from http://www.ed.gov/news/speeches/2009/11/11182009.html

Dunn, L. M., & Dunn, L. M. (1997). Peabody picture vocabulary test-revised. Circle Pines MN. *American Guidance Service.*

Early, D. M., Maxwell, K. L., Burchinal, M., Alva, S., Bender, R. H., Bryant, D., et al. (2007). Teachers' education, classroom quality, and young children's academic skills: Results from seven studies of preschool programs. *Child Development, 78*(2), 558-580.

Emmer, E. T., & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist, 36*(2), 103-112.

Entwisle, D. R., & Alexander, K. L. (1993). Entry into school: The beginning school transition and educational stratification in the United States. *Annual Review of Sociology, 19*, 401-423.

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *Journal of Educational Research, 94*(5), 275-282.

Feng, W. W., Jun, Y., & Xu, R. (2006). *A method/macro based on propensity score and Mahalanobis distance to reduce bias in treatment comparison in observational study*. Paper presented at the SAS Conference: PharmaSUG 2006. from http://www.lexjansen.com/pharmasug/2006/publichealthresearch/pr05.pdf

Frank Porter Graham Child Development Institute. (2003). Roadmaps to Quality. *Early Developments, 7*(2), 18-19.

Frank Porter Graham Child Development Institute. (2009). *Levels of training on the environment rating scales*. Retrieved from http://www.fpg.unc.edu/~ecers/training_levels.htm

Gates, B. (2009). *2009 annual letter*: Bill and Melinda Gates Foundation. http://www.gatesfoundation.org/annual-letter/Pages/2009-bill-gates-annual-letter.aspx

Goodson, B. D., Layzer, J. I., & Layzer, C. J. (2005). *Quality of early childhood care settings: Caregiver rating scale (QUEST)*. Cambridge, MA: Abt Associates Inc.

Halle, T., & Vick, J. (2007). *Quality in early childhood care and education settings: A compendium of measures*. Washington, DC: Child Trends for the Office of Planning, Research and Evaluation, Administration for Children and Families, US Department of Health and Human Services.

Hamre, B. K., LoCasale-Crouch, J., & Pianta, R. C. (2007). Formative assessment of classrooms: Using classroom observations to improve implementation quality. In L. M. Justice, C. Vukelich & W. H. Teale (Eds.), *Achieving Excellence in Preschool Literacy Instruction*: The Guilford Press.

Hamre, B. K., & Pianta, R. C. (2001). Early teacher-child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development, 72*(2), 625-638.

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*(5), 949-967.

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2008). Building a science of classrooms: Application of the CLASS framework in over 4,000 US early childhood and elementary classrooms. Retrieved December 1, 2008, from http://www.fcd-us.org/resources/resources_show.htm?doc_id=507559

Harms, T., & Clifford, R. (1989). *The Family Day Care Rating Scale*. New York: Teachers College Press.

Harms, T., & Clifford, R. M. (1980). *The Early Childhood Environment Rating Scale*. New York: Teachers College Press.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *The Early Childhood Environment Rating Scale: Revised edition*. New York: Teachers College Press.

Harms, T., Cryer, D., & Clifford, R. M. (1990). *Infant/Toddler Environment Rating Scale*. New York: Teachers College Press.

Harms, T., Jacobs, E., & White, D. (1996). *School Age Care Environment Rating Scale*. New York: Teachers College Press.

Harris, D. N., & Sass, T. R. (2006). Value-Added Models and the Measurement of Teacher Quality. *Preliminary Draft, Unpublished manuscript, Florida State University, April*.

Hightower, A. D., Work, W. C., Cowen, E. L., Lotyczewski, B. S., Spinell, A. P., Guare, J. C., et al. (1986). The Teacher-Child Rating Scale: A brief objective measure of elementary children's school problem behaviors and competencies. *School Psychology Review, 15*, 393-409.

Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review, 34*(4), 507.

Hintze, J. M., & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review, 33*(2), 258-271.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., et al. (2008). Ready to learn? Children's pre-academic achievement in pre-Kindergarten programs. *Early Childhood Research Quarterly, 23*(1), 27-50.

Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*(4), 403-424.

Hsieh, W. Y., Hemmeter, M. L., McCollum, J. A., & Ostrosky, M. M. (2009). Using coaching to increase preschool teachers' use of emergent literacy teaching strategies. *Early Childhood Research Quarterly*.

Jepsen, C. (2005). Teacher characteristics and student achievement: evidence from teacher surveys. *Journal of Urban Economics, 57*(2), 302-319.

Johnson, R. L., Penny, J. A., & Gordon, B. (2008). *Assessing performance: Designing, scoring, and validating performance tasks*. New York: The Guilford Press.

Justice, L. M., Pullen, P. C., Hall, A., & Pianta, R. C. (2003). *MyTeachingPartner language and literacy curriculum*. Charlottesville, VA: University of Virginia Center for Advanced Study of Teaching and Learning.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2007). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*.

La Paro, K. M., Hamre, B. K., LoCasale-Crouch, J., Pianta, R. C., Bryant, D. M., Early, D. M., et al. (in press). Quality in kindergarten classrooms: Observational evidence for the need to increase children's learning opportunities in early education classrooms. *Early Education & Development*.

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The Classroom Assessment Scoring System: Findings from the prekindergarten year. *Elementary School Journal, 104*(5), 409-426.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis, 24*(1), 37.

Leff, S. S., & Lakin, R. (2005). Playground-based observational systems: A review and implications for practitioners and researchers. *School Psychology Review, 34*(4), 474.

LoCasale-Crouch, J., Downer, J. T., & Hamre, B. K. (2010). *Assessing teacher beliefs about intentional instruction*: Manuscript in preparation.

LoCasale-Crouch, J., Konold, T., Pianta, R., Howes, C., Burchinal, M., Bryant, D., et al. (2007). Observed classroom quality profiles in state-funded pre-kindergarten programs and associations with teacher, program, and classroom characteristics. *Early Childhood Research Quarterly, 22*(1), 3-17.

Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2007). Does prekindergarten improve school preparation and performance? *Economics of Education Review, 26*(1), 33.

Magnuson, K. A., & Waldfogel, J. (2005). Early childhood care and education: Effects on ethnic and racial gaps in school readiness. *The Future of Children, 15*(1), 169-196.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O., Bryant, D., et al. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development, 79*(3), 732-749.

McCaffrey, D. F., Sass, T. R., & Lockwood, J. R. (2008). The intertemporal stability of teacher effect estimates. *Preliminary Draft, Unpublished manuscript, June*.

Medina, J. (2009, September 1). A 2-year study to learn what makes teachers good. *The New York Times*. Retrieved from http://cityroom.blogs.nytimes.com/2009/09/01/a-2-year-study-to-learn-what-makes-teachers-good/

Merrell, K. W. (1999). *Behavioral, social, and emotional assessment of children and adolescents*. Mahwah, N.J.: Lawrence Erlbaum.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics, 13*(2), 151-161.

Meyer, J. P., Henry, A. E., & Mashburn, A. J. (2009). *The influence of occasion on the reliability of classroom observations: An application of multivariate generalizability theory*. Paper presented at the Northeastern Educational Research Association, Rocky Mount, CT.

Michalopoulos, C., Bloom, H. S., & Hill, C. J. (2004). Can Propensity-Score Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs? *Review of Economics and Statistics, 86*(1), 156-179.

Montgomery, D. (2002). *Helping teachers develop through classroom observation*. London: David Fulton Publishers Ltd.

Moon, T. R., & Hughes, K. R. (2005). Training and scoring issues involved in large-scale writing assessments. *Educational Measurement: Issues and Practice, 21*(2), 15-19.

National Association for the Education of Young Children. (2009). Developmentally appropriate practice in early childhood programs serving children from birth through age 8. Retrieved February 24, 2009, from http://www.naeyc.org/about/positions.asp

National Association of Child Care Resource and Referral Agencies. (2009). *Comparison of Quality Rating and Improvement Systems (QRIS) with Department of Defense standards for quality* (No. 724-0714). Retrieved from http://www.naccrra.org/publications/naccrra-publications/comparison-qris-dod-quality-standards-2009.php

National Center for Education Statistics. (1994). *School and Staffing Survey 1993-1994; Principal's Survey*. Washington, DC: U.S. Department of Education.

National Child Care Information and Technical Assistance Center. (2007). *Child Care Bulletin Issue 32*. Fairfax, VA: Child Care Bureau, Administration for Children and Families, U.S. Department of Health and Human Services.

National Child Care Information and Technical Assistance Center. (2009). *Child care and development fund report of state and territory plans FY 2008-2009*. Retrieved from http://nccic.acf.hhs.gov/pubs/stateplan2008-09/index.html

National Council on Teacher Quality. (2007). State teacher policy yearbook: Progress on teacher quality. Retrieved January 1, 2008, from http://www.nctq.org/stpy/

Neuman, S. B., & Cunningham, L. (2009). The impact of professional development and coaching on early language and literacy instructional practices. *American Educational Research Journal, 46*(2), 532.

Neuman, S. B., Koh, S., & Dwyer, J. (2008). CHELLO: The Child/Home Environmental Language and Literacy Observation. *Early Childhood Research Quarterly, 23*(2), 159-172.

NICHD Early Child Care Research Network. (2002). The relation of global first-grade classroom environment to structural classroom features and teacher and student behaviors. *Elementary School Journal, 102*(5), 367-387.

NICHD Early Child Care Research Network. (2005). A day in third grade: A large-scale study of classroom quality and teacher and student behavior. *Elementary School Journal, 105*(3).

NICHD Early Child Care Research Network, & Duncan, G. J. (2003). Modeling the impacts of child care quality on children's preschool cognitive development. *Child Development, 74*(5), 1454-1475.

NICHD ECCRN. (2000). The relation of child care to cognitive and language development. *Child Development, 71*(4), 960-980.

NICHD ECCRN. (2002a). Child-Care Structure Process Outcome: Direct and Indirect Effects of Child-Care Quality on Young Children's Development. *Psychological Science, 13*(3), 199-206.

NICHD ECCRN. (2002b). The relation of global first-grade classroom environment to structural classroom features and teacher and student behaviors. *Elementary School Journal, 102*(5), 367-387.

NICHD ECCRN. (2005). A day in third grade: A large-scale study of classroom quality and teacher and student behavior. *Elementary School Journal, 105*(3).

NICHD ECCRN. (2006). Child-Care Effect Sizes for the NICHD Study of Early Child Care and Youth Development. *American Psychologist, 61*(2), 99-116.

NICHD ECCRN, & Duncan, G. J. (2003). Modeling the impacts of child care quality on children's preschool cognitive development. *Child Development, 74*(5), 1454-1475.

Norris, D. J., Dunn, L., & Eckert, L. (2003). *Reaching for the Stars: Center validation study final report*: Early Childhood Collaborative of Oklahoma.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large Are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78*(1), 33.

Peisner-Feinberg, E. S., Burchinal, M. R., Clifford, R. M., Culkin, M. L., Howes, C., Kagan, S. L., et al. (2001). The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child Development, 72*(5), 1534-1553.

Penny, J. A. (2003). Reading high stakes writing samples: My life as a reader. *Assessing Writing, 8*(3), 192-215.

Pianta, R. C. (2001). Student Teacher Relationship Scale. Lutz, FL: Psychological Assessment Resources, Inc.

Pianta, R. C., & Allen, J. P. (2008). Building capacity for positive youth development in secondary school classrooms: Changing teachers' interactions with students. In *Toward Positive Youth Development: Transforming Schools and Community Programs*.

Pianta, R. C., Belsky, J., Houts, R., Morrison, F., & NICHD ECCRN. (2007). Observed classroom experiences in elementary school: A day in fifth grade and stability from grades 1 to 3. Manuscript submitted for publication.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109.

Pianta, R. C., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., et al. (2005). Features of pre-kindergarten programs, classrooms, and teachers: Do they predict observed classroom quality and child-teacher interactions? *Applied Developmental Science, 9*(3), 144-159.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008a). *Classroom Assessment Scoring System*. Baltimore, MD: Brookes Publishing.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008b). Technical appendix. In *Classroom Assessment Scoring System* (pp. 87-106). Baltimore, MD: Brookes Publishing.

Pianta, R. C., La Paro, K. M., Payne, C., Cox, M. J., & Bradley, R. (2002). The relation of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes. *The Elementary School Journal, 102*(3), 225(215).

Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher–child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly*.

Preschool Curriculum Evaluation Research Consortium. (2008). *Effects of preschool curriculum programs on school readiness* (NCER No. 2008-2009). Washington, DC: Institute of Education Sciences, National Center for Education Research. Retrieved from http://ncer.ed.gov

Pressley, M., Wharton-McDonald, R., Allington, R., Collins Block, C., Morrow, L., Tracey, D., et al. (2001). A study of effective first-grade literacy instruction. *Scientific Studies of Reading, 5*(1), 35-58.

Ramey, C. T., Campbell, F. A., Burchinal, M., Skinner, M. L., Gardner, D. M., & Ramey, S. L. (2000). Persistent effects of early childhood education on high-risk children and their mothers. *Applied Developmental Science, 4*(1), 2-14.

Raudenbush, S. W. (2005). Learning from Attempts to Improve Schooling: The Contribution of Methodological Diversity. *Educational Researcher, 34*(5), 25.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. London: Sage Publications.

Raudenbush, S. W., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness*, *1*(2), 138-154.

Raudenbush, S. W., & Sampson, R. J. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology, 29*, 1-41.

Reynolds, A. J., Temple, J. A., Robertson, D. L., & Mann, E. A. (2001). Long-term effects of an early childhood intervention on educational achievement and juvenile arrest: A 15-year follow-up of low-income children in public schools. *JAMA: Journal of the American Medical Association, 285*(18), 2339-2346.

Rimm-Kaufman, S. E., Curby, T. W., Grimm, K. J., Nathanson, L., & Brock, L. L. (2009). The contribution of children's self-regulation and classroom quality to children's adaptive behaviors in the kindergarten classroom. *Developmental Psychology, 45*(4), 958-972.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417-458.

Robelen, E. W. (2008, November 12). Gates' new approach gets good reviews. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2008/11/12/13gatesreact.h28.html

Robelen, E. W. (2009, January 22). Gates gives $22 million in grants. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2009/01/22/19gates.h28.html

Rock, D. A., & Stenner, A. J. (2005). Assessment issues in the testing of children at school entry. *The Future of Children, 15*(1), 15(20).

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247-252.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33-38.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*(3-4), 169-188.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95*(450), 573-585.

Schaefer, E. S., & Edgerton, M. (1985). Parent and child correlates of parental modernity. *Parental belief systems: The psychological consequences for children,* 287–318.

Schafer, W. D., Gagné, P., & Lissitz, R. W. (2005). Resistance to Confounding Style and Content in Scoring Constructed-Response Items. *Educational Measurement: Issues and Practice, 24*(2), 22-28.

Schweinhart, L. J., & Weikart, D. P. (1997). The High/Scope Preschool Curriculum Comparison Study through age 23. *Early Childhood Research Quarterly, 12*(2), 117.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can randomized experiments yield accurate answers?  A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association, 103*(484), 1334-1356.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*: Houghton Mifflin.

Simon, A., & Boyer, E. G. (1969). *Mirrors for Behavior, An Anthology of Classroom Observation Instruments*. Philadelphia, PA: Research for Better Schools, Inc.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23*(4), 323.

Smith, J. (2000). *A critical survey of empirical methods for evaluating active labor market policies*: Department of Economics, University of Western Ontario.

Smith, M., Dickinson, D. K., Sangeorge, A., & Anastasopoulos, L. (2002). Early language and literacy classroom observation (ELLCO) toolkit: Baltimore: Brookes.

Stipek, D., & Byler, P. (2004). The early childhood classroom observation measure. *Early Childhood Research Quarterly, 19*(3), 375-397.

Stoney, L. (2004). *Financing Quality Rating Systems: Lessons learned*: Alliance for Early Childhood Finance for United Way of America Success by 6.

Stuhlman, M., Curby, T. W., Grimm, K. J., Mashburn, A., Chomat-Mooney, L., Hamre, B. K., et al. (2009). Within-day variability in third and fifth grade in classroom interaction quality: Implications for children's experience and conducting classroom observation. *Manuscript in preparation*.

Stuhlman, M., & Pianta, R. C. (2009). Profiles of educational quality in first grade. *The Elementary School Journal, 109*(4), 323-342.

Teachstone. (2009). *CLASS regional training*. Retrieved from http://www.teachstone.org/regional_training.php

Tout, K., Starr, R., & Cleveland, J. (2008). *Evaluation of Parent Aware: Minnesota's Quality Rating System pilot*. Minneapolis, MN: Minnesota Early Learning Foundation Research Consortium.

Tout, K., Zaslow, M., Halle, T., & Forry, N. (2009). *Issues for the next decade of Quality Rating and Improvement Systems* (Issue Brief No. 3). Washington, DC: Prepared by Child Trends for the Office of Planning, Research, and Evaluation, Administration for Children and Families, US Department of Health and Human Services.

U.S. Department of Health and Human Services, Administration for Children and Families, & Office of Head Start. (2008). *Classroom Assessment Scoring System* (Information Memorandum No. ACF--IM-HS-08-11). Washington, DC: Brown, Patricia E.

Wachs, T. D., Gurkas, P., & Kontos, S. (2004). Predictors of preschool children's compliance behavior in early childhood classroom settings. *Journal of Applied Developmental Psychology, 25*(4), 439-457.

Wasik, B. A., Bond, M. A., & Hindman, A. (2006). The effects of a language and literacy intervention on Head Start children and teachers. *Journal of Educational Psychology, 98*(1), 63.

Wiley, C., Good, T., & McCaslin, M. (2008). Comprehensive school reform instructional practices throughout a school year: The role of subject matter, grade level, and time of year. *The Teachers College Record, 110*(11), 2361-2388.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III: Tests of Achievement*. Itasca, ILL: Riverside Publishing.

Zaslow, M., Tout, K., Halle, T., & Forry, N. (2009). *Multiple purposes for measuring quality in early childhood settings: Implications for collecting and communicating information on quality* (Issue Brief No. 2). Washington, DC: Prepared by Child Trends for the Office of Planning, Research, and Evaluation, Administration for Children and Families, US Department of Health and Human Services.

Zellman, G. L., & Perlman, M. (2008). *Child-care Quality Rating and Improvement Systems in five pioneer states: Implementation issues and lessons learned*. Arlington, VA: RAND Corporation.

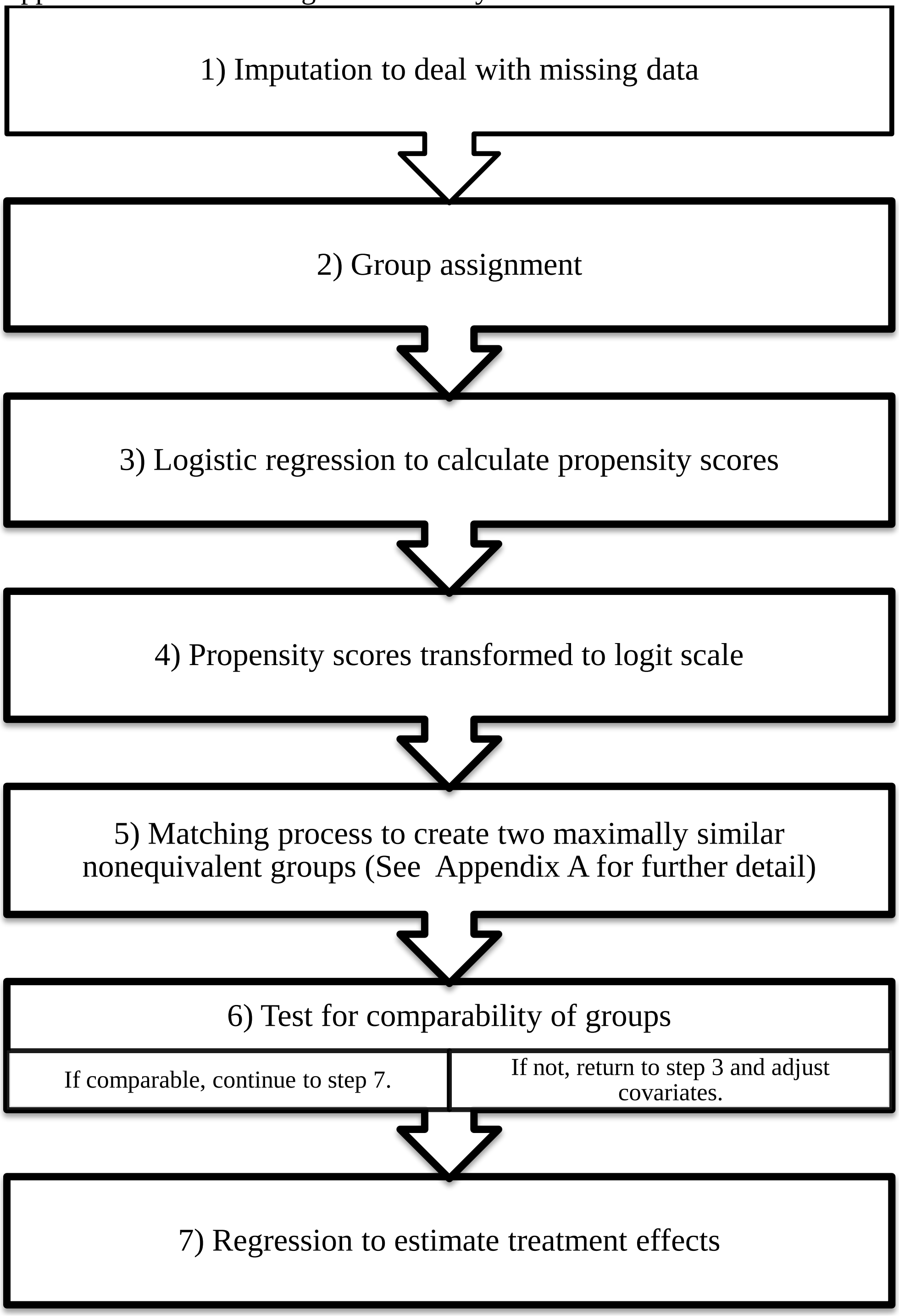
Zellman, G. L., Perlman, M., Le, V.-N., & Setodji, C. M. (2008). *Assessing the validity of the Qualistar Early Learning Quality Rating and Improvement System as a tool for improving child-care quality*. Arlington, VA: RAND Corporation.

Zepeda, S. J. (2008). *The instructional leader's guide to informal classroom observations*. Larchmont, NY: Eye on Education.

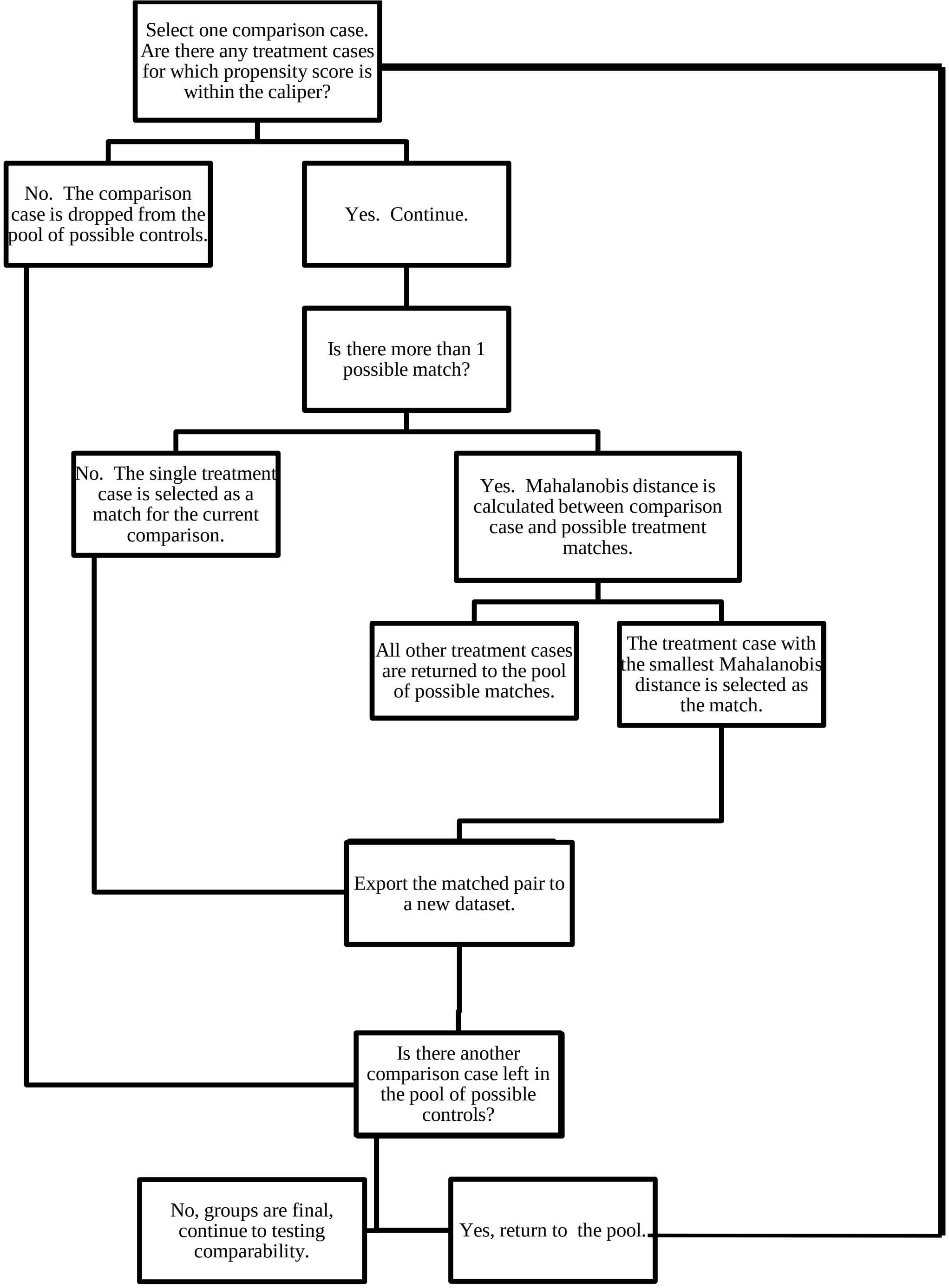
Zill, N., Resnick, G., Kim, K., McKey, R. H., Clark, C., Pai-Samant, S., et al. (2001). *Head Start FACES: Longitudinal Findings on Program Performance. Third Progress Report.* http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED453969&site=ehost-live

Zill, N., Sorongon, A., Kim, K., Clark, C., & Woolverton, M. (2006). *FACES 2003 research brief*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families. Retrieved from www.acf.hhs.gov/programs/opre/hs/faces/index.html

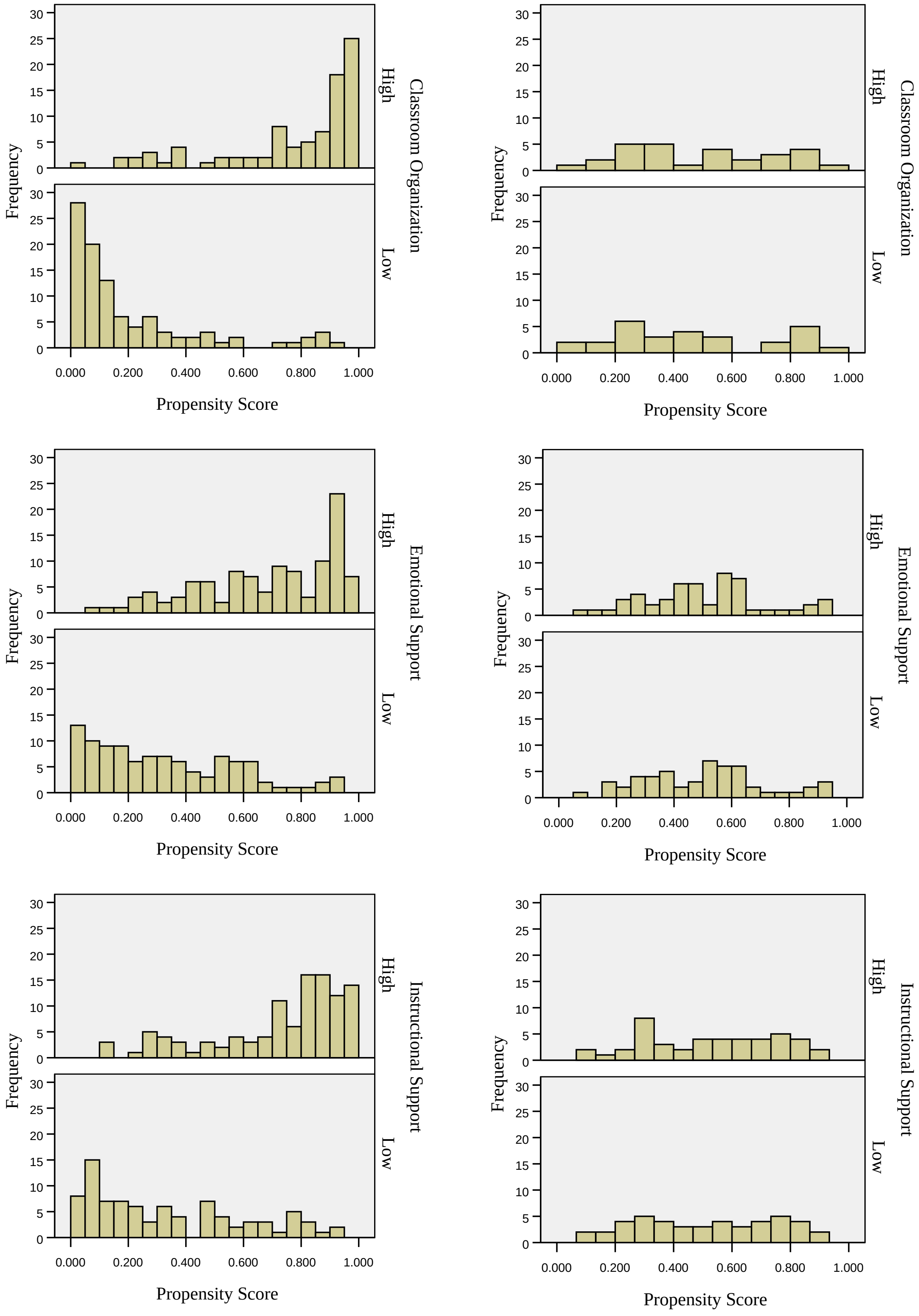
Appendix A. Overall design of data analysis for each of the CLASS domains.


1) Imputation to deal with missing data
2) Group assignment
3) Logistic regression to calculate propensity scores
4) Propensity scores transformed to logit scale
5) Matching process to create two maximally similar nonequivalent groups (See  Appendix A for further detail)
6) Test for comparability of groups
If comparable, continue to step 7.
If not, return to step 3 and adjust covariates.
7) Regression to estimate treatment effects

Appendix B. Matching process.


Select one comparison case. Are there any treatment cases for which propensity score is within the caliper?
No. The comparison case is dropped from the pool of possible controls.
Yes. Continue.
Is there more than 1 possible match?
No. The single treatment case is selected as a match for the current comparison.
Yes. Mahalanobis distance is calculated between comparison case and possible treatment matches.
All other treatment cases are returned to the pool of possible matches.
The treatment case with the smallest Mahalanobis distance is selected as the match.
Export the matched pair to a new dataset.
Is there another comparison case left in the pool of possible controls?
No, groups are final, continue to testing comparability.
Yes, return to the pool.

Appendix C. Propensity score distribution for each CLASS domain, pre- and post-match.


Pre-Match

Post-Match

Footnote

[1] Standardized regression weights were calculated by dividing the raw regression coefficients by each outcome's known standard deviation (from the standardization sample), and multiplying by the standard deviation of the Instructional Quality measure. Our calculations should be interpreted with caution, because level-1 and level-2 variances from an unconditional model were not published and thus were not incorporated in our calculations.

Table 1

*CLASS Cutoff (Mean) Scores for Group Assignment*

| Group | Classroom Organization | Emotional Support | Instructional Support |
|---|---|---|---|
| Low | | | |
| Pre-K | 4.05 (3.40) | 4.79 (4.24) | 1.73 (1.40) |
| K | 4.35 (3.90) | 5.11 (4.54) | 1.64 (1.36) |
| Pre-Match *n* | 103 | 109 | 94 |
| High | | | |
| Pre-K | 4.83 (5.38) | 5.54 (5.92) | 2.42 (3.08) |
| K | 4.98 (5.35) | 5.65 (6.00) | 2.06 (2.65) |
| Pre-Match *n* | 89 | 111 | 113 |

Table 2

*Assessing Bias Reduction and Group Comparability*

| Comparison | Organization | Emotional | Instructional |
|---|---|---|---|
| Pre-match | | | |
| *n* | 187 | 211 | 195 |
| Difference in mean propensity scores | 1.42 | 1.18 | 1.23 |
| Ratio of propensity score variances | 0.68 | 1.18 | 1.03 |
| % covariates with variance ratio | | | |
| ≤1/2 | 0.12 | 0.06 | 0.12 |
| >1/2 and ≤4/5 | 0.12 | 0.18 | 0.06 |
| >4/5 and ≤5/4 | 0.59 | 0.47 | 0.65 |
| >5/4 and ≤2 | 0.06 | 0.24 | 0.06 |
| >2 | 0.12 | 0.06 | 0.12 |
| Post-match | | | |
| *n* | 56 | 106 | 90 |
| Difference in mean propensity scores | 0.07 | 0.02 | 0.06 |
| Ratio of propensity score variances | 0.98 | 1.03 | 1.10 |
| % covariates with variance ratio | | | |
| ≤1/2 | 0.00 | 0.00 | 0.00 |
| >1/2 and ≤4/5 | 0.18 | 0.06 | 0.24 |
| >4/5 and ≤5/4 | 0.65 | 0.94 | 0.71 |
| >5/4 and ≤2 | 0.18 | 0.00 | 0.06 |
| >2 | 0.00 | 0.00 | 0.00 |

Table 3

*Effects of Teacher-Child Interactions on Kindergarten Spring Outcomes*

| | Organization, *n*=56 | | | | Emotional, *n*=106 | | | | Instructional, *n*=90 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Outcome | *B* | *p* | *SE* | *d* | *B* | *p* | *SE* | *d* | *B* | *p* | *SE* | *d* |
| Direct Assessment | | | | | | | | | | | | |
| Letter Naming | 2.36 | * | 1.02 | 0.59 | 0.22 | | 0.98 | 0.05 | 0.06 | | 0.71 | 0.01 |
| Number Naming | 0.50 | | 0.35 | 0.41 | 0.29 | | 0.32 | 0.24 | 0.04 | | 0.18 | 0.03 |
| Counting 1-1 | -3.29 | | 3.08 | -0.33 | -0.40 | | 2.04 | -0.04 | -1.67 | | 2.09 | -0.17 |
| PPVT (SS) | 8.79 | ** | 3.06 | 0.75 | 3.81 | | 2.38 | 0.33 | 4.05 | | 2.56 | 0.35 |
| OWLS (SS) | 3.39 | | 2.88 | 0.27 | 1.47 | | 2.40 | 0.12 | 4.33 | † | 2.49 | 0.35 |
| Letter Word Id (SS) | 6.14 | † | 3.14 | 0.49 | 1.99 | | 2.71 | 0.16 | 2.37 | | 2.13 | 0.19 |
| Applied Problem (SS) | 2.29 | | 3.07 | 0.20 | -0.69 | | 2.34 | -0.06 | 0.54 | | 2.44 | 0.05 |
| Rhyming (raw) | — | | — | — | 1.00 | | 0.86 | 0.22 | 2.07 | * | 0.92 | 0.45 |
| Teacher Report | | | | | | | | | | | | |
| Social Competence | 0.23 | | 0.21 | 0.28 | 0.00 | | 0.16 | 0.00 | -0.12 | | 0.16 | -0.15 |
| Behavior Problems | -0.22 | | 0.15 | -0.31 | -0.09 | | 0.13 | -0.13 | 0.12 | | 0.12 | 0.18 |
| ARS Language | 0.62 | * | 0.25 | 0.63 | — | | — | — | 0.20 | | 0.18 | 0.20 |
| ARS Math | 0.46 | † | 0.25 | 0.48 | 0.19 | | 0.19 | 0.20 | 0.21 | | 0.19 | 0.22 |
| STRS Closeness | 0.19 | | 0.19 | 0.29 | 0.10 | | 0.13 | 0.15 | 0.03 | | 0.17 | 0.04 |
| STRS Conflict | -0.12 | | 0.18 | -0.15 | -0.25 | † | 0.13 | -0.32 | 0.12 | | 0.15 | 0.15 |

*Note.* PPVT = Peabody Picture Vocabulary Test; OWLS = Oral and Written Language Scale; ARS = Academic Rating Scale; STRS = Teacher-child Relationship Scale; SS = standard score.

†*p* < .10. **p* < .05. ***p*< .01.

# Accounting for Variance in an Observational Measure of Teacher Quality

Anne E. Henry

Robert C. Pianta

University of Virginia

## Abstract

Observational assessment is being used at a large scale to evaluate the quality of interactions between teachers and children in classroom environments. One of the issues for evaluators who use observational assessment across large numbers of classrooms is that the decisions made regarding observation protocol can introduce error to observed scores, limiting power to test intervention effects or introducing bias to estimates of the relationships between observed scores and outcomes. This study looks at the relationship between observed scores using the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, and Hamre, 2008) and characteristics of the protocol, specifically the day of the week, month of the year, and duration of an observation cycle in minutes. Results indicate that the timing of CLASS observations accounts for very little variation in observed scores, with a few exceptions for the CLASS domain of Classroom Organization. This suggests that teachers' interactions with children are fairly stable across these protocol factors.

# Accounting for Variance in an Observational Measure of Teacher Quality

There has been intensified pressure on policymakers and school administrators in recent years to identify teachers who are most effective in improving children's outcomes and to support teachers who are less effective. In kindergarten through 12th grade settings, this pressure is related to the authorization of No Child Left Behind (NCLB) and associated changes to teacher licensing systems at the state level. In early childhood education, this pressure is related to the rapid development of Quality Rating Systems (QRS) at the state level (Barnett et al., 2008; National Child Care Information and Technical Assistance Center, 2009). Central challenges to the accurate identification of effective teachers include pinpointing characteristics and practices of teachers that are predictive of children's outcomes and then finding ways to measure these characteristics and practices at a large scale.

Observation is one approach that has been used for assessing and improving teacher quality, including observation of teachers' interactions with children and their implementation of curricula (e.g. Barnett et al., 2008; Connor, Piasta et al., 2009; Davidson, Fields, & Yang, 2009; Mashburn et al., 2008). Various observational measures have been used successfully to describe differences in practices between teachers and programs (Bryant, Clifford, & Peisner, 1991; NICHD ECCRN, 2002a; Pianta et al., 2005; Pianta et al., 2002). Scores from these observations have also been shown to predict children's academic and social skills (Burchinal et al., 2008; Burchinal et al., 2009; Hamre & Pianta, 2005; Mashburn et al., 2008). Moreover, there is evidence of classroom observation being used for both evaluation and professional development at a large scale. Scores from observations of teacher-child interactions are increasingly used in an evaluative way; for example, 38 states require site visits to monitor the quality of state-funded pre-kindergarten (Barnett et al., 2008); observational measures are often involved as part of their Quality Rating Systems (National Child Care Information and Technical Assistance Center, 2007, 2009). Also, of $22 million in grants recently distributed by the Bill & Melinda Gates Foundation and designated for research on teacher effectiveness, a portion of the funding will be used to videotape and observe teachers (Medina, 2009; Robelen, 2008, 2009), putting some of the best examples online “as a model for other teachers and as a resource for students” (Gates, 2009, p. 12).

One of the challenges for evaluators who use observational assessment at a large scale is that the decisions made regarding observation protocol can influence the reliability and validity of the data collected and how that data should be interpreted. There is typically a great deal of variation in observed scores, both between teachers (Bryant et al., 1991; NICHD ECCRN, 2002a; Pianta et al., 2005; Pianta et al., 2002) and between observation occasions for the same teacher (Curby, Brock et al., 2009; J. P. Meyer et al., 2009; Pianta, La Paro et al., 2008a; Zellman & Perlman, 2008). Variation in observed scores may be explained by teacher, child, or protocol characteristics if these things are also carefully measured. For example, teacher:child ratios, activity types, or time of day observed may account for variance in the quality of observed interactions among teachers and children if measured appropriately. Otherwise, unaccounted-for variance is labeled as measurement “error” and can become problematic in large-scale studies of teachers, classrooms, and student outcomes. When measurement error is present, statistical power to estimate the impact of quality improvement interventions is reduced and bias is introduced to estimates of the association between quality and children's outcomes (Raudenbush & Sadoff, 2008).

Still, potential sources of variation in observed scores such as activity setting or the day of the week observed can be controlled for through careful planning of the observation protocol and/or during later analyses. Depending on the questions of interest, collection of observational data can be limited to certain days of the week, instructional activities, group sizes, etc. Alternatively, observers can note these variables in addition to assigning ratings and evaluators can control for them in later data analysis.

Since the number of variables that can be measured is limited by available resources in terms of time and money, research is needed to describe the influence of each of these potential sources of variance so evaluators and principal investigators of large-scale studies can make informed decisions about which is most important to control for.

The purpose of the current study is to evaluate the influence of a few potential sources of variance in ratings of observed teacher-child interactions that evaluators often have control over when planning a protocol or analyzing data - characteristics of the protocol related to the timing and duration of observations. Specifically, does the day of the week the observation occurs, month of the year the observation occurs, or duration of the observation account for within-teacher variability in observed scores? We begin by describing the advantages and challenges of using observational tools for measuring classroom processes.

## Advantages of Classroom Observation for Assessing Teacher-Child Interaction Quality

Observation of classroom environments can be an effective way to measure the quality of teachers' and children's experiences in those environments (Pianta & Hamre, 2009), and can be used to describe the nature of interactions among teachers and children (e.g. Connor, Morrison et al., 2009; Hamre et al., 2008; Wiley, Good, & McCaslin, 2008) or the implementation of an intervention (O'Donnell, 2008; Raudenbush, 2005). There are two advantages to classroom observation over other methodologies for measuring teacher quality. First, observed scores of teacher-child interactions are predictive of children's academic, social, and behavioral outcomes (Connor, Piasta et al., 2009; Mashburn et al., 2008; NICHD ECCRN, 2002a; Wasik, Bond, & Hindman, 2006), whereas the associations between more commonly used proxies for teacher effectiveness, such as certification or teacher education, and children's outcomes are mixed or nonexistent (Boyd, Goldhaber, Lankford, & Wyckoff, 2007; Clotfelter, Ladd, & Vigdor, 2007; Early et al., 2007; Jepsen, 2005; Kane, Rockoff, & Staiger, 2007). Second, ratings are based on specific behaviors of teachers and students that can be objectively defined and developed in professional development contexts (Pianta & Hamre, 2009). This is an improvement over using calculations of teachers' "value-added" to students' achievement to identify effective teachers (e.g. Nye et al., 2004; Rivkin et al., 2005) because the same standards used to evaluate teachers can also effectively be integrated into professional development through coursework or consultancy models to increase teachers' knowledge and practice of effective interactions with children (Dickinson & Caswell, 2007; Hsieh, Hemmeter, McCollum, & Ostrosky, 2009; Neuman & Cunningham, 2009; Pianta, Mashburn, Downer, Hamre, & Justice, 2008).

Large-scale evaluation and professional development via classroom observation requires the development of standardized measures and protocols to facilitate interpretation. Many standardized measures are available with demonstrated reliability and validity; particularly measures designed for observing early childhood education quality (i.e. Danielson, 1996; Harms et al., 1998; Neuman, Koh, & Dwyer, 2008; Pianta, La Paro et al., 2008a; M. Smith, Dickinson, Sangeorge, & Anastasopoulos, 2002). For example, the Early Childhood Environment Rating Scale (ECERS; Harms & Clifford, 1980) and its later revision (ECERS-R; Harms et al., 1998) have been widely used to assess child care quality. For the ECERS-R, 43 items are categorized into seven subscales, and each item is scored on a 7-point scale from 1 (inadequate) to 7 (excellent). The subscales assess whether a classroom has appropriate routines, activities, and materials in place for children, provisions for parents and staff, and whether teaching staff interact with children in developmentally appropriate ways. The Early Language and Literacy Classroom Observation (ELLCO) Toolkit (M. Smith et al., 2002) is another observation instrument that has been widely used in early childhood settings. The ELLCO is comprised of three tools, a literacy environment checklist, an observation and teacher interview, and a literacy activity rating scale. Observers using these three tools of the ELLCO collect information about the materials available and a teacher’s approach to facilitating children’s language and literacy skills.

In recent years, research has drawn attention to the quality of observed interactions among teachers and children as being particularly important for children’s developmental outcomes (Burchinal et al., 2008; Burchinal et al., 2009; Connor et al., 2005; Curby, LoCasale-Crouch et al., 2009; Hamre & Pianta, 2005; Howes et al., 2008;

Mashburn et al., 2008; NICHD ECCRN, 2005; Pianta et al., 2002; Rimm-Kaufman et al., 2009). In fact, measures of children's direct experiences in classrooms appear more predictive of their outcomes than measures of the structural features of classrooms such as teacher qualifications or program location (Howes et al., 2008). For example, in classrooms where teachers were observed to interact with children in respective and responsive ways, the children demonstrated greater levels of social competence (Burchinal et al., 2009; Mashburn et al., 2008), greater growth in phonological awareness (Curby, LoCasale-Crouch et al., 2009), and higher vocabulary and decoding scores (Connor et al., 2005). Children in classrooms that received higher scores on observational scales of the instructional support provided by teachers demonstrated higher levels of academic and language skills (Burchinal et al., 2008; Howes et al., 2008; Mashburn et al., 2008). Children whose teachers were observed to have established clear routines and proactive approaches to discipline have greater levels of behavioral and cognitive self-control, spend less time off-task (Rimm-Kaufman et al., 2009), and greater gains in math skills (Curby, LoCasale-Crouch et al., 2009).

One standardized observation tool that has been used to assess the quality of teacher-child interactions is the Classroom Assessment Scoring System, or CLASS (Pianta, La Paro et al., 2008a). Observed scores from the CLASS and its predecessor, the Classroom Observation System (COS; NICHD ECCRN, 2002a), predict children's academic, social, and behavioral outcomes (Burchinal et al., 2008; Howes et al., 2008; Mashburn et al., 2008; NICHD ECCRN, 2005; NICHD ECCRN & Duncan, 2003). The CLASS is increasingly being used at a large scale, as part of early childhood Quality Rating Systems in Minnesota and Virginia (Tout et al., 2009) and nationwide monitoring and quality improvement for the Office of Head Start (U.S. Department of Health and Human Services et al., 2008). Secretary of Education Arne Duncan also drew attention to the CLASS in recent remarks at the National Association for the Education of Young Children annual conference (Duncan, 2009).

## Challenges to Using Observational Methodology and Possible Responses

Although authors of standardized observational tools typically provide some recommendations for conducting observations, there are few empirically-based guidelines for observation protocols known to increase reliability and decrease measurement error when observation is conducted at a large scale (Raudenbush & Sadoff, 2008). Evaluators are faced with the challenge to make decisions on observation protocols that lead to the most efficient use of resources while still maintaining or improving upon the predictive validity of the instrument.

When planning large-scale observational assessment, evaluators should first look to what information is already available. As observation of classroom environments is used more frequently to assess and improve teacher quality, there is evidence of tremendous variation in observed quality between teachers (Bryant et al., 1991; NICHD ECCRN, 2002a; Pianta et al., 2005; Pianta et al., 2002). Interestingly, when the same teacher is observed by multiple raters or over multiple occasions, there is even more variability in the observed scores assigned to a single teacher (Curby, Brock et al., 2009; J. P. Meyer et al., 2009; Pianta, La Paro et al., 2008a; Zellman & Perlman, 2008). Within-teacher variability can be present whether one rater observes the same teacher on multiple occasions, or multiple raters observe the same teacher on a single occasion. For example, observers using the CLASS for live coding are advised to begin coding at the start of the school day, use 30-minute cycles (20-minute observe, 10-minute score), and obtain a minimum of four cycles (Pianta, La Paro et al., 2008a, p. 10).

Within-teacher variability is important for evaluators using observation at a large scale to be aware of for two reasons, further discussed below. First, within-teacher variation could represent error in measurement, and thus impact estimates of the association between observed quality and children's outcomes (Raudenbush & Sadoff, 2008). Second, this variation could be substantively meaningful – the materials or teacher-child interactions present in a classroom could vary over time and this could be important in predicting children's outcomes. Research is needed to better understand potential sources of within-teacher variation so evaluators can make informed decisions when planning observation protocol, reacting to data collection challenges, and interpreting findings.

To clarify, one reason within-teacher variability is important is because it could reflect measurement error. When classroom observations are conducted as part of a large-scale research study, measurement error reduces statistical power to assess the effects of interventions to improve observed teacher-child interactions, and introduces bias into estimates of the association between teacher-child interactions and children's outcomes (Raudenbush & Sadoff, 2008). Multiple sources of measurement error are possible when using observational tools, such as rater effects, or the day of the week or month of the year of observation (Pianta & Hamre, 2009; Raudenbush & Sadoff, 2008; Raudenbush & Sampson, 1999). For example, to assess the quality of teacher-child interactions, multiple raters may observe a classroom on multiple days. On each day, each rater may observe and assign scores for the same time period, perhaps repeating for multiple observation occasions within a single day. The observed score representing the quality of teacher-child interactions is calculated by aggregating the scores across raters, days, and occasions. However, this aggregated score may be influenced by the day of observation, time of observation, or the raters who assigned scores.

There are many possible reasons why variability in observed scores for the same teacher could be substantively meaningful. Some of this variance could be providing systematic information about children, teachers, and classrooms. Differences in activity settings (i.e. math, reading, transitions), group size (i.e. large group, small group), or time of day observed (i.e. morning, after lunch) could influence the presence or absence of observed quality indicators. As an example, there may be many opportunities for teachers to engage children in higher-order thinking skills like prediction during a science activity, but fewer opportunities as children wash their hands before snack-time. Characteristics of the observation protocol, such as rater, day of the week observed, or the duration of the observation occasion, could also be substantively meaningful in explaining within-teacher variation. Children (or raters!) may be less focused late on a Friday afternoon and observed scores could reflect that.

There has been some research to examine temporal stability of observed scores and how this may influence children’s outcomes. One study reports increasing negativity and chaos in third and fifth grade classrooms over the school day, and lower quality of instructionally supportive interactions in the first 30 minutes of the school day, suggesting that students and teachers take some time to “settle in” to instruction early in the day, and may be subject to fatigue as the day continues (Stuhlman et al., 2009). In another study, observed instruction in the fall was more structured and focused on basic skills than instruction in the spring, potentially reflecting adjustments based on students' ability levels over the course of the year (Wiley et al., 2008). Variability could also be relevant for children's academic and behavioral outcomes; in one study the consistency of emotionally supportive teacher-child interactions within a day was more predictive of children's outcomes than the mean level of emotional support (Curby, Brock et al., 2009).

Knowing more about how these potential sources of variation influence observed scores becomes really useful when evaluators are faced with the logistics of assessing teacher quality at a large scale. Most education researchers are well aware of the real-world challenges to classroom observation protocols – unscheduled recess or fire drills, early school closures limit data collection, rescheduling due to sick days, poor camera angles on videotaped observations, etc. Questions come up in response to these challenges, such as whether observed ratings from the month of December are roughly equivalent estimates of quality to those from October or April, or the percentage of observations which must be double-coded. In some cases, decisions must be made to include or exclude scores from observations after data collection is complete, with little information regarding which issues are the most problematic.

Other methodologies for assessing teacher quality are not immune to these challenges regarding within-teacher variability. For example, instability in estimates of teacher effects from year to year using value-added modeling may be related to the structure of the model, non-random assignment of students to teachers, or lack of reliability in achievement measures (Harris & Sass, 2006; McCaffrey, Sass, & Lockwood, 2008). Given the demonstrated relationship between observed teacher-child interactions and children's outcomes (Burchinal et al., 2008; Burchinal et al., 2009; Curby, LoCasale-Crouch et al., 2009; Mashburn et al., 2008) and the additional usefulness of observational tools for professional development (e.g. Dickinson & Caswell, 2007; Neuman & Cunningham, 2009; Pianta, Mashburn et al., 2008), further work to address the challenges of observational assessment is important. By studying potential sources of variance in observational methodology more explicitly, we can better understand the most important sources of variance and either restrict or measure them in later studies.

## Dealing with Multiple Sources of Variation in Observational Assessment

In many cases, when scores vary across observations of the same teacher, evaluators will aggregate the scores to estimate children's overall exposure to classroom experiences over a given time period. Combining scores in this way facilitates comparison between teachers. For example, this is common when using scores from the CLASS. The CLASS manual recommends averaging cycle-level scores across the total number of completed cycles to create teacher-level composite CLASS scores (Pianta, La Paro et al., 2008a, p. 17). Some researchers have instead chosen to randomly select one cycle per teacher for use in analyses (e.g. Stuhlman & Pianta, 2009). These techniques are supported by evidence that CLASS scores are pretty stable across cycles in a day and across days in a week, with domain and dimension-level correlations being moderate to high (Pianta, La Paro et al., 2008a, pp. 96-99). Scores are less stable across months of the year, with only low to moderate correlations between dimensions in the fall and spring (Pianta, La Paro et al., 2008a, p. 100). Less-than-perfect stability in CLASS scores across multiple observation cycles indicates that there is some variability in scores occurring between cycles within observations of the same teacher and children.

The creation of composite scores when multiple observation cycles of each teacher are available allows us to make comparisons between teachers (Bryant et al., 1991; NICHD ECCRN, 2002a; Pianta et al., 2005; Pianta et al., 2002), but hides within-teacher variability in CLASS domain and dimension scores. There are a few approaches to study sources of within-teacher variability more explicitly.

One analytic approach that has been used to understand multiple sources of variation present in observed scores is Generalizability theory (G theory; Cronbach et al., 1972). G theory was originally designed to assess the reliability and validity of individual-level measures, but can also be adapted for setting-level assessment (Hintze, 2005; Raudenbush & Sampson, 1999). Using G theory, reliability and error variance are evaluated relative to the context of observation. Evaluators can estimate the portions of error variance that can be accounted for by appropriate situational variables (e.g. raters, time of day, and day of week). Evaluators can then fine-tune observation protocols based on what they have learned regarding major sources of measurement error. Moreover, G studies can be used in conjunction with Decision (D) studies through which evaluators can assess the degree of reduction in measurement error following manipulation of the observation protocol (Brennan, 2001; Hintze, 2005).

Although G theory can be useful for identifying which situational variables are the largest sources of variance in observed scores, it tells evaluators little about how situational variables influence scores. As described above, some variation may also be giving us important information about teachers and children in classrooms. Knowing whether certain situational variables raise or lower observed scores may influence the questions evaluators ask and how data is interpreted. For example, if evaluators are working on an intervention to maximize teachers' use of instructional time and minimize time spent on managerial activities, they may be interested to learn that scores from morning observations reflect a high number of managerial activities which then decreases over the course of the day.

Understanding how the observation context influences scores can be accomplished through a simpler analytic approach. The direction and degree to which situational variables influence observed scores can be assessed using regression coefficients. In the current study, characteristics of observation protocol that evaluators often have control over will be examined to see if they influence scores of the observed quality of teacher-child interactions.

### The Current Study

The purpose of the current paper is to evaluate the extent to which the timing of observation cycles accounts for within-teacher variability in observed CLASS scores. Hierarchical Linear Modeling (Raudenbush & Bryk, 2002) will be used to assess whether the day of the week of the observation cycle, the month of the year of the cycle, or the duration of observation account for a significant portion of the variability in CLASS scores within teachers who were observed for multiple cycles.

## Method

### Participants

Participants in this study included 56 prekindergarten teachers and represented a subset of 239 teachers participating in a larger study, MyTeachingPartner (MTP), of the impacts of professional development supports on teacher and child outcomes. The study took place in state-funded preschools in a mid-Atlantic state. Teachers participating in MTP were randomly assigned to conditions varying in level of professional development support. For the current study, 56 teachers were included based on two criteria, they: 1) were one of 91 teachers randomly assigned to the "Web-Only" condition, and 2) participated in the project over the full two years of the study. Teachers exposed to this condition of MTP received a low level of professional development support. Support included access to web-based lesson plans from the MTP-Language and Literacy (MTP-LL; Justice, Pullen, Hall, & Pianta, 2003) curriculum and the Promoting Alternative Thinking Strategies (PATHS; Domitrovich, Greenberg, Kusche, & Cortes, 2004) curriculum in social competence; teachers also received access to the MTP website which provided 1-2 minute video exemplars of high quality teacher-child interactions and text describing the interactions using the language of the CLASS framework. Teachers in the Web-Only condition were asked to implement MTP-LL activities for ten minutes per day and PATHS activities once a week.

Of the 56 teachers included in the current paper, all but one teacher had a bachelor's degree at the time of the study. Twenty (36%) of teachers also had an advanced degree. Teachers had an average of 10.7 years (SD = 8.5) of experience working professionally with preschool-aged children, with a range from 0 to 33 total years of experience. Teachers' were on average 45 years old (SD = 9.3, range 25 – 61 years).

Children in classrooms taught by study teachers were enrolled in a state-funded pre-kindergarten program targeted to serve a population of children meeting risk indicators for early school difficulties, including having a family income below Federal poverty guidelines; family stressors such as homelessness, unemployment, low levels of parent/guardian education, or chronic illness; developmental delays; or limited English proficiency. On average, 66% of students in each study classroom had families with income below Federal poverty guidelines; this varied at the classroom level from 14 to 100%. Classrooms, on average, were 51% male and 19% White, and had 15 children enrolled (range 6 to 19 children).

## Measures

**Observed teacher-child interactions.** The Classroom Assessment Scoring System (CLASS; Pianta, La Paro et al., 2008a) was used to assess the quality of teacher-child interactions. The CLASS has been validated through standardized observations in more than 4,000 prekindergarten and elementary classrooms (Hamre et al., 2008). The tool was developed through careful literature review and feedback from professionals in the fields of psychology and education. Observed scores from the CLASS and its predecessor, the Classroom Observation System (COS; NICHD ECCRN, 2002a), predict children's academic, social, and behavioral outcomes (Burchinal et al., 2008; Howes et al., 2008; Mashburn et al., 2008; NICHD ECCRN, 2005; NICHD ECCRN & Duncan, 2003). Observers assigned global ratings on a 7-point scale to each of nine dimensions of teacher-child interactions; scores of 1-2 are low-range scores, and 6-7 are high-range. For each dimension, observers looked for specific teacher and child behaviors, richly described in the manual and known to be important for children's development, and scores were based on both the consistency and quality of these behaviors. The nine dimensions are organized into three domains of support available to students in classrooms, further described below: Emotional Support, Instructional Support, and Classroom Organization.

The domain of *Emotional Support* is comprised of three dimensions, including Positive Climate, Negative Climate (reversed), and Teacher Sensitivity. Positive Climate reflects warmth, respect, and emotional connections among the teacher and students as communicated through verbal and nonverbal interactions. Negative Climate indicates the frequency, quality, and intensity of teacher and peer negativity. Teacher Sensitivity reflects a teacher's awareness of and responsiveness to students' academic and emotional needs. A fourth dimension, Regard for Student Perspectives, which captures the extent to which a teacher emphasizes students' interests and autonomy in the classroom, was revised during this study and thus excluded from the present analyses.

The domain of *Instructional Support* is comprised of three dimensions, including Concept Development, Quality of Feedback, and Language Modeling. Concept Development reflects a teacher's focus on developing students' higher-order thinking skills and understanding of concepts. Quality of Feedback describes the extent to which a teacher provides students with feedback that expands their learning and participation. Language Modeling indicates a teacher's use of techniques to model and facilitate language for students.

The domain of *Classroom Organization* is comprised of three dimensions, including Behavior Management, Productivity, and Instructional Learning Formats. Behavior Management indicates a teacher's use of effective methods to prevent and redirect misbehavior. Productivity reflects a teacher's effective management of instructional time and provision of learning opportunities. Instructional Learning Formats captures a teacher's facilitation of student engagement and learning.

Observers learned to assign CLASS scores by attending two days of training followed by a calibration test of three videos. Before they were allowed to code, observers scored three videos and were required to score within one point of master scores for each dimension on at least 80% of all scores. Observer reliability was further supported through weekly meetings when all observers watched and discussed a video together.

Inter-rater reliability was computed for a subset of CLASS scores. Two observers assigned scores for 33 randomly selected tapes and were considered to be in agreement if their scores were within one point of each other. Agreement ranged from 79% for the dimension Instructional Learning Formats to 97% for Productivity. These rates of agreement are comparable to those seen during live observation in large-scale studies using the CLASS (La Paro, Pianta, & Stuhlman, 2004; NICHD ECCRN, 2002b, 2005).

**Classroom characteristics.** Information about classroom demographics was gathered from two sources and used as covariates in the present analyses. A classroom is defined both in terms of teacher and study cohort. Teachers in the present analyses participated in the study for two years and have classroom characteristic data for both cohort 1 and cohort 2. The first source of classroom characteristic data was survey completed by the teacher at the beginning of each year. Two items were used for this paper – the number of children enrolled in the class, and the percentage of children who are male. The second source of information was a survey completed by parents/guardians of all children in each classroom. Information about family income was aggregated to the classroom level.

## Procedures

Teachers were asked to send in 30-minute videos of their teaching every two weeks for the duration of the project, labeled with the date the video was recorded. They were asked to implement an activity for each video, alternating between PATHS and MTP-LL activities. They were also given guidelines about when to tape - a few minutes prior to the start of the activity, through the full activity, and continuing after the activity was completed up to at least 30 minutes. Teachers sent in an average of 11 videos each (SD = 4.9), with some teachers sending as few as 1 video or as many as 21 videos.

Videos were coded using the CLASS, but were excluded from the present study if they were shorter than 5 minutes long. Videos were 24.9 minutes long on average (SD = 7.13). Several conditions were in place to guide the amount of time observed before scores were assigned. Observers started coding when all of the following conditions were met: the classroom was visible on the video, audio was present, and children and/or the teacher were present. Coding stopped when any one of the following conditions was met: the video stopped, 30 minutes had passed, or at least one activity was complete and the children and teacher were off-screen for longer than 5 minutes. As observers coded, they made note of the time-related parameters of interest in this paper: the duration of the observation in minutes and the date the video was recorded. From the date of observation, the authors were able to discern the month of the year that was videotaped as well as the day of the week.

## Analysis

The data for this study were hierarchically nested, with multiple observations nested within each teacher. To assess the associations between timing of observation cycles and CLASS scores while controlling for nesting of observations within teachers, we conducted two-level hierarchical linear modeling (HLM; Raudenbush & Bryk, 2002) using SAS Proc Mixed (Singer, 1998). The outcomes of interest were CLASS domain scores for Emotional Support, Instructional Support, and Classroom Organization. The domain scores for each observation were computed by averaging the appropriate dimension scores (i.e. scores for Positive Climate, Reversed Negative Climate, and Teacher Sensitivity were averaged to create an Emotional Support domain score). Analyses were conducted separately for each of the three CLASS domains.

We were interested in the associations between these domain scores and three characteristics of observation cycles (level-1): length of observation in minutes, day of the week the video was recorded, and month of the year the video was recorded. The length of observation was centered at the mean length (24.9 minutes). The day of the week a video was recorded was entered as four dummy-coded variables to represent Monday through Friday. Videos recorded on Mondays were used for the reference group. Likewise, the month of the year a video was recorded was entered as 11 dummy-coded variables and September was used as the reference group. Although the study took place over two years, each variable representing a month was for the two years combined. For example, October 2004 and October 2005 were combined to create a single variable representing October. The decision to use Monday and September as reference groups was supported by data indicating that teachers who sent in tapes that were recorded on Mondays or in the month of September were not significantly different from teachers who did not, in terms of years of experience, possessing an advanced degree, age, or their beliefs about children. All of these predictors were fixed at level 1, as we did not expect the association between characteristics of observation cycles and CLASS domain scores to vary between teachers. See Table 1 for the frequency of videos that were sent in for given days of the week and months of the year.

Additional covariates were also included as fixed effects at the observation level (level-1) to control for characteristics of classrooms that may influence the association between timing of observation cycles and CLASS domain scores. This was important given that two years of observations were included in the analyses, and each teacher was observed with two cohorts of children. We included four such characteristics: the percentage of children in a classroom whose families had income below the federal poverty level, the percentage of children who were male, the number of children enrolled in the classroom at the beginning of the year, and cohort. Also, the total number of videos per teacher was included as a covariate at level 2 (teacher).

In the final model, the CLASS domain score for observation i in teacher j includes the intercept or overall domain score, plus the contributions of day of the week, month of the year, duration of the observation in minutes, as well as additional level-1 covariates including cohort, number of students, percentage of students who were male, and percentage of students who were poor. The intercept is further defined at level-2 by the total number of videos per teacher, plus the error for teacher j. Intercept differences were allowed to vary across teachers, but the effects of $\beta_{1-17}$ were fixed at the observation-level.

$Y_{ij} = \beta_{0j} + \beta_{1\text{-}4}$(day of the week video was recorded) + $\beta_{5\text{-}12}$(month of the year video was recorded) + $\beta_{13}$(duration of observation) + $\beta_{14}$(cohort) + $\beta_{15}$(number of students) + $\beta_{16}$(percentage male) + $\beta_{17}$(percentage poor) + $r_{ij}$

$\beta_{0j} = \gamma_{00} + \gamma_{01}$(total number of videos per teacher) + $u_0$

$\beta_{1\text{-}17j} = \gamma_{1\text{-}17}$

**Results**

We first established that there was substantial within-teacher variance in an unconditional model. The intraclass correlation, ρ, for Emotional Support was .29, for Classroom Organization it was .21, and for Instructional Support it was .12. There was more than twice as much variation in CLASS domain scores within teachers observed on multiple occasions as there was between teachers.  In other words, teacher-child interactions varied quite a bit from one observation to the next, even when the same teacher was being observed.

Intercepts, unstandardized coefficients (β), and standard errors (SE) are presented in Table 2 for each of the three outcomes. Day of observation, month of observation, and length in minutes of observation generally did not account for significant variance at level 1, within-teachers, for the CLASS domains of Emotional Support or Instructional Support, with a single exception. There was a trend toward a higher level of Instructional Support in the month of January relative to the month of September.  For the most part, variation in teachers' warmth and responsiveness toward children was not related to the timing of observation cycles in terms of day, month, or duration, and neither was variation in teachers' attempts to engage children in higher-order thinking skills.

In Classroom Organization, however, two observation characteristics accounted for significant within-teacher variation. The quality of Classroom Organization was significantly higher when observations were conducted in the month of February and significantly lower when classrooms were observed for more minutes. Teachers used more effective and consistent methods to prevent and redirect misbehavior and organize instructional time during observation cycles videotaped in February than they did during September observations. Also, methods for organizing instructional time were less effective as classrooms were observed for longer periods.

**Follow-up analyses.** When we found significant predictors for the domain of Classroom Organization, we wanted to see if the associations were due to idiosyncrasies at the CLASS dimension level. Recall that each CLASS domain is the average of several dimension scores. Classroom Organization is the average of scores in Behavior Management, Productivity, and Instructional Learning Formats. Unconditional models indicated that there was significant variance at level 1 for each of these dimensions; the intraclass correlations, ρ, were .23, .13, and .19, respectively.

We then added the same predictors and covariates as before. Results for all three dimensions appear in Table 3. There was a significant, positive association between the quality of Behavior Management and observations occurring in February, relative to September. There was evidence of less misbehavior and more proactive management midway in the year relative to the start of the year. Length of observation in minutes was negatively associated with scores in Productivity. Teachers were less consistent in effectively managing instructional time and reducing time for managerial tasks as observers watched for longer periods. None of the predictors were significant for

Instructional Learning Formats, meaning that teachers' facilitation of children's engagement was not influenced by the day of the week, month of the year, or duration of the observation.

Interestingly, additional predictors were significant for the dimension of Productivity. Namely, the associations between Productivity scores and observations occurring on Tuesdays, Wednesdays, or in March were significant and positive. There were more learning opportunities present and fewer managerial tasks conducted during observations that occurred mid-week relative to at the start or end of the week.

## Discussion

This study confirmed that there is significant variation in teacher-child interactions from one observation occasion to the next, but this variation is generally not related to the day of the week, month of the year, or duration of the observation occasion. This is particularly true when using CLASS *domain* scores of Emotional Support, Instructional Support, and Classroom Organization. While there is still plenty of research to be done to examine sources of within-teacher variation in observed scores, evaluators can move forward with some confidence that CLASS domain scores would not be affected if observations were to occur on one day/month or another, or if classrooms are observed for 15 minutes versus 25. However, there are some instances of timing significantly influencing CLASS *dimension* scores (e.g. Behavior Management, Productivity, Instructional Learning Formats). Since dimension scores are typically averaged to create the CLASS domain scores, this does not matter in most cases. But if evaluators have questions that are dimension-specific, some caution is warranted.

More specifically, there were just a few significant associations between CLASS scores and the day of the week of observation, month of observation, or length of observation. The significant associations that did appear are possible to explain conceptually. At the domain level, it is possible that the quality of Classroom Organization is higher in February relative to September because teachers need time at the beginning of the year to establish clear classroom rules and routines, and children need time to learn them. There is evidence to suggest that this shift in organization over the course of the year can be good for children's outcomes; first graders' whose teachers were observed to spend more time on organization in the fall had stronger letter and word reading skills in the spring (Cameron, Connor, Morrison, Jewkes, 2008).

The negative association between Classroom Organization and length of observation in minutes can also be easily interpreted. When observations are longer, there are more opportunities for observers to see children misbehave. In addition, teachers who sent in videotapes that were shorter in length may not have videotaped much time following activities, whereas teachers who sent in longer tapes may have included transitions between activities in the video.  Unless teachers establish clear expectations for students' behavior during transitions, keep them brief, and embed learning opportunities within them, transitions can negatively affect Classroom Organization scores.

These interpretations of the associations were supported by further exploration at the dimension level. The coefficient for length of observation is negative and significant for Productivity, fitting with the theory that longer observations offer more opportunities for seeing ineffective transitions between activities. Though not significant, the coefficient for length of observation is also negative for Behavior Management, suggesting that there are also more opportunities to observe misbehavior. Coefficients indicating the association between the month of observation and Productivity, though not significant in all of the spring months, are all positive, suggesting that teachers are able to better manage instructional time once routines are established in the fall. None of the predictors were associated with Instructional Learning Formats, or teacher's facilitation of students' engagement.

**Implications.** The findings of this study indicate that CLASS scores are generally resilient to fluctuations in the day of week, month of year, and duration of observation occasions. Considering the significant associations that were found, CLASS dimension scores appear slightly more sensitive to the timing of observations. For this reason, evaluators are advised to primarily use CLASS domain-level scores (Emotional Support, Instructional Support, and Classroom Organization) for data analysis. Also, evaluators who are planning observation protocol should ensure that data is collected across all of the dimensions so that the creation of domain composites is possible. Although some evaluators may be tempted to select dimensions specific to their interests, the domain scores are less sensitive to observation timing and evaluators can be more confident in their results using scores at that level.

If evaluators are truly interested in specific aspects of teacher-child interactions at the CLASS dimension level, they should be thoughtful of variability that could be substantively meaningful. In the case of Behavior Management and Productivity, dimensions in the domain of Classroom Organization, changes in scores from one month to another, or as observations take place over longer periods, could simply reflect rhythms and routines that take place in classrooms over time.

By studying how characteristics of protocol influence observed CLASS scores, we gain insight on whether variance in CLASS scores should be considered error or systematic variability. Significant sources of variation should be measured intentionally, so that they can be described and controlled for when appropriate. This will only improve our ability to identify change in teacher-child interactions over the course of interventions, and give power to estimates of the effect of teacher-child interactions on children's outcomes.

# References

Barnett, W. S., Epstein, D. J., Friedman, A. H., Boyd, J. S., & Hustedt, J. T. (2008). *The state of preschool 2008: State preschool yearbook*. New Brunswick, NJ: National Institute for Early Education Research, Rutgers University.

Boyd, D., Goldhaber, D., Lankford, H., & Wyckoff, J. (2007). The Effect of Certification and Preparation on Teacher Quality. *Future of Children, 17*(1), 45-68.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Bryant, D. M., Clifford, R. M., & Peisner, E. S. (1991). Best Practices for Beginners: Developmental Appropriateness in Kindergarten. *American Educational Research Journal, 28*(4), 783.

Burchinal, M., Howes, C., Pianta, R. C., Bryant, D., Early, D. M., Clifford, R. M., et al. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher–child interactions and instruction. *Applied Developmental Science, 12*(3), 140-153.

Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2009). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly*.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review, 26*(6), 673-682.

Connor, C. M. D., Morrison, F. J., Fishman, B. J., Ponitz, C. C., Glasney, S., Underwood, P. S., et al. (2009). The ISI Classroom Observation System: Examining the Literacy Instruction Provided to Individual Students. *Educational Researcher, 38*(2), 85.

Connor, C. M. D., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., et al. (2009). Individualizing student instruction precisely: Effects of child by instruction interactions on first graders' literacy development. *Child Development, 80*(1), 77–100.

Connor, C. M. D., Son, S. H., Hindman, A. H., & Morrison, F. J. (2005). Teacher qualifications, classroom practices, family characteristics, and preschool experience: Complex effects on first graders' vocabulary and early reading outcomes. *Journal of School Psychology, 43*(4), 343-375.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Curby, T. W., Brock, L. L., & Hamre, B. K. (2009). The role of consistency in preschool teacher-child interactions. *Manuscript under review*.

Curby, T. W., LoCasale-Crouch, J., Konold, T. R., Pianta, R. C., Howes, C., Burchinal, M., et al. (2009). The Relations of Observed Pre-K Classroom Quality Profiles to Children's Achievement and Social Competence. *Early Education & Development, 20*(2), 346-372.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Davidson, M. R., Fields, M. K., & Yang, J. (2009). A randomized trial study of a preschool literacy curriculum: The importance of implementation. *Journal of Research on Educational Effectiveness, 2*(3), 177-208.

Dickinson, D. K., & Caswell, L. (2007). Building support for language and early literacy in preschool classrooms through in-service professional development: Effects of the Literacy Environment Enrichment Program (LEEP). *Early Childhood Research Quarterly, 22*(2), 243-260.

Domitrovich, C. E., Greenberg, m. T., Kusche, C., & Cortes, R. (2004). *The preschool PATHS curriculum*. State College, PA: Pennsylvania State University.

Duncan, A. (2009). *The early learning challenge: Raising the bar*. Retrieved from http://www.ed.gov/news/speeches/2009/11/11182009.html

Early, D. M., Maxwell, K. L., Burchinal, M., Alva, S., Bender, R. H., Bryant, D., et al. (2007). Teachers' education, classroom quality, and young children's academic skills: Results from seven studies of preschool programs. *Child Development, 78*(2), 558-580.

Gates, B. (2009). *2009 annual letter*: Bill and Melinda Gates Foundation. http://www.gatesfoundation.org/annual-letter/Pages/2009-bill-gates-annual-letter.aspx

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*(5), 949-967.

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2008). Building a science of classrooms: Application of the CLASS framework in over 4,000 US early childhood and elementary classrooms. Retrieved December 1, 2008, from http://www.fcd-us.org/resources/resources_show.htm?doc_id=507559

Harms, T., & Clifford, R. M. (1980). *The Early Childhood Environment Rating Scale*. New York: Teachers College Press.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *The Early Childhood Environment Rating Scale: Revised edition*. New York: Teachers College Press.

Harris, D. N., & Sass, T. R. (2006). Value-Added Models and the Measurement of Teacher Quality. *Preliminary Draft, Unpublished manuscript, Florida State University, April*.

Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review, 34*(4), 507.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., et al. (2008). Ready to learn? Children's pre-academic achievement in pre-Kindergarten programs. *Early Childhood Research Quarterly, 23*(1), 27-50.

Hsieh, W. Y., Hemmeter, M. L., McCollum, J. A., & Ostrosky, M. M. (2009). Using coaching to increase preschool teachers' use of emergent literacy teaching strategies. *Early Childhood Research Quarterly*.

Jepsen, C. (2005). Teacher characteristics and student achievement: evidence from teacher surveys. *Journal of Urban Economics, 57*(2), 302-319.

Justice, L. M., Pullen, P. C., Hall, A., & Pianta, R. C. (2003). *MyTeachingPartner language and literacy curriculum*. Charlottesville, VA: University of Virginia Center for Advanced Study of Teaching and Learning.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2007). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*.

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The Classroom Assessment Scoring System: Findings from the prekindergarten year. *Elementary School Journal, 104*(5), 409-426.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O., Bryant, D., et al. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development, 79*(3), 732-749.

McCaffrey, D. F., Sass, T. R., & Lockwood, J. R. (2008). The intertemporal stability of teacher effect estimates. *Preliminary Draft, Unpublished manuscript, June*.

Medina, J. (2009, September 1). A 2-year study to learn what makes teachers good. *The New York Times*. Retrieved from http://cityroom.blogs.nytimes.com/2009/09/01/a-2-year-study-to-learn-what-makes-teachers-good/

Meyer, J. P., Henry, A. E., & Mashburn, A. J. (2009). *The influence of occasion on the reliability of classroom observations: An application of multivariate generalizability theory*. Paper presented at the Northeastern Educational Research Association, Rocky Mount, CT.

National Child Care Information and Technical Assistance Center. (2007). *Child Care Bulletin Issue 32*. Fairfax, VA: Child Care Bureau, Administration for Children and Families, U.S. Department of Health and Human Services.

National Child Care Information and Technical Assistance Center. (2009). *Child care and development fund report of state and territory plans FY 2008-2009*. Retrieved from http://nccic.acf.hhs.gov/pubs/stateplan2008-09/index.html

Neuman, S. B., & Cunningham, L. (2009). The impact of professional development and coaching on early language and literacy instructional practices. *American Educational Research Journal, 46*(2), 532.

Neuman, S. B., Koh, S., & Dwyer, J. (2008). CHELLO: The Child/Home Environmental Language and Literacy Observation. *Early Childhood Research Quarterly, 23*(2), 159-172.

NICHD ECCRN. (2002a). Child-Care Structure Process Outcome: Direct and Indirect Effects of Child-Care Quality on Young Children's Development. *Psychological Science, 13*(3), 199-206.

NICHD ECCRN. (2002b). The relation of global first-grade classroom environment to structural classroom features and teacher and student behaviors. *Elementary School Journal, 102*(5), 367-387.

NICHD ECCRN. (2005). A day in third grade: A large-scale study of classroom quality and teacher and student behavior. *Elementary School Journal, 105*(3).

NICHD ECCRN, & Duncan, G. J. (2003). Modeling the impacts of child care quality on children's preschool cognitive development. *Child Development*, *74*(5), 1454-1475.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large Are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78*(1), 33.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109.

Pianta, R. C., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., et al. (2005). Features of pre-kindergarten programs, classrooms, and teachers: Do they predict observed classroom quality and child-teacher interactions? *Applied Developmental Science, 9*(3), 144-159.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System*. Baltimore, MD: Brookes Publishing.

Pianta, R. C., La Paro, K. M., Payne, C., Cox, M. J., & Bradley, R. (2002). The relation of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes. *The Elementary School Journal, 102*(3), 225(215).

Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher–child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly*.

Raudenbush, S. W. (2005). Learning from Attempts to Improve Schooling: The Contribution of Methodological Diversity. *Educational Researcher, 34*(5), 25.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. London: Sage Publications.

Raudenbush, S. W., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness, 1*(2), 138-154.

Raudenbush, S. W., & Sampson, R. J. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology, 29*, 1-41.

Rimm-Kaufman, S. E., Curby, T. W., Grimm, K. J., Nathanson, L., & Brock, L. L. (2009). The contribution of children's self-regulation and classroom quality to children's adaptive behaviors in the kindergarten classroom. *Developmental Psychology, 45*(4), 958-972.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Robelen, E. W. (2008, November 12). Gates' new approach gets good reviews. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2008/11/12/13gatesreact.h28.html

Robelen, E. W. (2009, January 22). Gates gives $22 million in grants. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2009/01/22/19gates.h28.html

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23*(4), 323.

Smith, M., Dickinson, D. K., Sangeorge, A., & Anastasopoulos, L. (2002). Early language and literacy classroom observation (ELLCO) toolkit: Baltimore: Brookes.

Stuhlman, M., Curby, T. W., Grimm, K. J., Mashburn, A., Chomat-Mooney, L., Hamre, B. K., et al. (2009). Within-day variability in third and fifth grade in classroom interaction quality: Implications for children's experience and conducting classroom observation. *Manuscript in preparation*.

Stuhlman, M., & Pianta, R. C. (2009). Profiles of educational quality in first grade. *The Elementary School Journal*, *109*(4), 323-342.

Tout, K., Zaslow, M., Halle, T., & Forry, N. (2009). *Issues for the next decade of Quality Rating and Improvement Systems* (Issue Brief No. 3). Washington, DC: Prepared by Child Trends for the Office of Planning, Research, and Evaluation, Administration for Children and Families, US Department of Health and Human Services.

U.S. Department of Health and Human Services, Administration for Children and Families, & Office of Head Start. (2008). *Classroom Assessment Scoring System* (Information Memorandum No. ACF--IM-HS-08-11). Washington, DC: Brown, Patricia E.

Wasik, B. A., Bond, M. A., & Hindman, A. (2006). The effects of a language and literacy intervention on Head Start children and teachers. *Journal of Educational Psychology, 98*(1), 63.

Wiley, C., Good, T., & McCaslin, M. (2008). Comprehensive school reform instructional practices throughout a school year: The role of subject matter, grade level, and time of year. *The Teachers College Record, 110*(11), 2361-2388.

Zellman, G. L., & Perlman, M. (2008). *Child-care Quality Rating and Improvement Systems in five pioneer states: Implementation issues and lessons learned*. Arlington, VA: RAND Corporation.

Table 1
*Frequency of Videotapes Recorded on a Given Day or in a Given Month*

| Time | Number |
|---|---|
| Day of week | |
| Monday | 98 |
| Tuesday | 88 |
| Wednesday | 125 |
| Thursday | 171 |
| Friday | 131 |
| Total | 613 |
| Month of year | |
| September | 57 |
| October | 106 |
| November | 91 |
| December | 50 |
| January | 48 |
| February | 78 |
| March | 61 |
| April | 63 |
| May | 59 |
| Total | 613 |

Table 2

*Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for CLASS Domains*

| | Emotional Support | | | | Instructional Support | | | | Classroom Organization | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Parameter | Model 1 | | Model 2 | | Model 1 | | Model 2 | | Model 1 | | Model 2 | |
| | | | | | Fixed effects | | | | | | | |
| Intercept | 5.15 *** | (0.08) | 5.40 *** | (0.53) | 2.79 *** | (0.06) | 2.43 *** | (0.49) | 5.20 *** | (0.06) | 5.05 *** | (0.46) |
| Level-1 (observations) | | | | | | | | | | | | |
| Day of week | | | | | | | | | | | | |
| Tuesday | | | 0.08 | (0.12) | | | 0.08 | (0.14) | | | 0.11 | (0.11) |
| Wednesday | | | 0.14 | (0.12) | | | 0.06 | (0.13) | | | 0.15 | (0.11) |
| Thursday | | | 0.00 | (0.11) | | | 0.00 | (0.12) | | | -0.06 | (0.10) |
| Friday | | | 0.06 | (0.11) | | | -0.11 | (0.13) | | | -0.03 | (0.11) |
| Month of year | | | | | | | | | | | | |
| October | | | 0.01 | (0.14) | | | 0.04 | (0.15) | | | 0.05 | (0.13) |
| November | | | -0.21 | (0.14) | | | -0.05 | (0.16) | | | -0.16 | (0.13) |
| December | | | 0.12 | (0.16) | | | 0.16 | (0.18) | | | 0.20 | (0.15) |
| January | | | 0.19 | (0.17) | | | 0.36 † | (0.19) | | | 0.12 | (0.16) |
| February | | | 0.06 | (0.15) | | | 0.14 | (0.17) | | | 0.31 * | (0.14) |
| March | | | -0.05 | (0.15) | | | -0.03 | (0.17) | | | 0.21 | (0.14) |
| April | | | -0.18 | (0.15) | | | 0.04 | (0.17) | | | 0.14 | (0.14) |
| May | | | -0.25 | (0.16) | | | -0.06 | (0.18) | | | 0.02 | (0.15) |
| Length (minutes) | | | 0.00 | (0.01) | | | 0.01 | (0.01) | | | -0.01 * | (0.01) |
| | | | | | Random effects | | | | | | | |
| Level-2 Intercept (uj) | 0.26 *** | (0.06) | 0.28 *** | (0.07) | 0.11 ** | (0.04) | 0.11 ** | (0.04) | 0.14 *** | (0.04) | 0.15 *** | (0.04) |
| Level-1 Residual (rij) | 0.63 *** | (0.04) | 0.61 *** | (0.04) | 0.82 *** | (0.05) | 0.78 *** | (0.05) | 0.54 *** | (0.03) | 0.54 *** | (0.03) |

*Note*. Standard errors are in parentheses. Models also included controls for the percentage of children in a classroom whose families had income below the federal poverty level, the percentage of children who were male, the number of children enrolled in the classroom at the beginning of the year, cohort, and the number of observations per teacher.

†*p* <.10. **p* <.05. ***p* <.01. ****p* <.001.

Table 3

*Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for CLASS Dimensions*

| | Behavior Management | | | | Productivity | | | | Instructional Learning Formats | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Parameter | Model 1 | | Model 2 | | Model 1 | | Model 2 | | Model 1 | | Model 2 | |
| | | | | | Fixed effects | | | | | | | |
| Intercept | 5.25 *** | (0.09) | 5.42 *** | (0.68) | 5.64 *** | (0.06) | 5.55 *** | (0.50) | 4.71 *** | (0.08) | 4.29 *** | (0.60) |
| Level-1 (observations) | | | | | | | | | | | | |
| Day of week | | | | | | | | | | | | |
| Tuesday | | | 0.04 | (0.17) | | | 0.35 * | (0.14) | | | -0.03 | (0.15) |
| Wednesday | | | 0.03 | (0.16) | | | 0.28 * | (0.14) | | | 0.16 | (0.15) |
| Thursday | | | -0.18 | (0.15) | | | 0.03 | (0.13) | | | 0.00 | (0.14) |
| Friday | | | -0.09 | (0.16) | | | 0.08 | (0.13) | | | -0.06 | (0.14) |
| Month of year | | | | | | | | | | | | |
| October | | | -0.01 | (0.19) | | | 0.06 | (0.16) | | | 0.12 | (0.17) |
| November | | | -0.08 | (0.19) | | | -0.19 | (0.16) | | | -0.19 | (0.18) |
| December | | | 0.39 † | (0.22) | | | 0.20 | (0.19) | | | 0.02 | (0.20) |
| January | | | 0.18 | (0.23) | | | 0.10 | (0.20) | | | 0.08 | (0.21) |
| February | | | 0.44 * | (0.20) | | | 0.33 † | (0.17) | | | 0.17 | (0.19) |
| March | | | 0.15 | (0.21) | | | 0.41 * | (0.18) | | | 0.08 | (0.19) |
| April | | | 0.12 | (0.21) | | | 0.24 | (0.18) | | | 0.09 | (0.19) |
| May | | | 0.23 | (0.21) | | | 0.00 | (0.18) | | | -0.15 | (0.20) |
| Length (minutes) | | | -0.01 | (0.01) | | | -0.03 *** | (0.01) | | | 0.00 | (0.01) |
| | | | | | Random effects | | | | | | | |
| Level-2 Intercept (uj) | 0.36 *** | (0.09) | 0.36 *** | (0.09) | 0.13 ** | (0.05) | 0.10 ** | (0.04) | 0.22 *** | (0.06) | 0.23 *** | (0.07) |
| Level-1 Residual (rij) | 1.16 *** | (0.07) | 1.15 *** | (0.07) | 0.88 *** | (0.05) | 0.86 *** | (0.05) | 0.93 *** | (0.06) | 0.97 *** | (0.06) |

*Note*. Standard errors are in parentheses. Models also included controls for the percentage of children in a classroom whose families had income below the federal poverty level, the percentage of children who were male, the number of children enrolled in the classroom at the beginning of the year, cohort, and the number of observations per teacher.

†*p* <.10. **p* <.05. ***p* <.01. ****p* <.001.

# Rater Calibration When Observational Assessment Occurs at Large-Scale: Degree of Calibration and Characteristics of Raters Associated with Calibration

Anne E. Henry

Bridget K. Hamre

Robert C. Pianta

University of Virginia

# Abstract

Observational assessment is being used to study the quality of interactions among teachers and children across large numbers of classrooms, but training a workforce of raters that can assign reliable scores when observations are used in large-scale contexts can be challenging and expensive. Two issues for evaluators include training large numbers of raters to calibrate to an observation tool and identifying raters who are capable of calibration. This study looks at the extent of rater calibration across 2,093 raters trained by the Office of Head Start (OHS) in 2008-2009, and for a subsample of 704 raters, characteristics that predict their calibration. Findings indicate that it is possible to train large numbers of raters to calibrate to an observation tool and rater beliefs about teachers and children predicted the degree of calibration. Implications for large-scale observational assessments are discussed.

## Rater Calibration When Observational Assessment Occurs at Large-Scale: Degree of Calibration and Characteristics of Raters Associated with Calibration

Direct observation can be used to study children and teachers in school settings and is increasingly used at a large scale to assess and improve teacher quality and effectiveness (Pianta & Hamre, 2009). There are many observation systems available to researchers, school administrators, school psychologists, and teachers. These systems vary widely in their degree of evidence-based reliability and validity and in the level of training provided to people interested in using them, ranging from principal-developed observation strategies used in case studies of individual teachers (e.g. Colvin, Flannery, Sugai, & Monegan, 2009; Montgomery, 2002) to researcher-developed observation systems complete with commercially-available manuals, trainings, and scoresheets (e.g. Harms et al., 1998; Pianta, La Paro et al., 2008a; M. Smith et al., 2002). Although some of these tools have been studied extensively, and developers have collected data considered both reliable and valid, the best strategies for passing standardized, evidence-based tools on to other researchers, administrators, and teachers, and training them to collect data that is also reliable and valid, are still unclear. The Office of Head Start (OHS) recently faced this challenge when adopting an observational tool for the purpose of monitoring grantees. The current paper uses the OHS efforts as an example, to understand the degree to which raters were successfully able to calibrate to an observation tool when trained at large-scale and the characteristics of raters that are associated with their calibration.

Information on best practices for observational assessment is much needed as direct observation is increasingly being used to describe and evaluate teacher performance and classroom quality. Observation measures were included in several large-scale research studies of early childhood education settings, including the National Institute of Child Health and Human Development (NICHD) Study of Early Child Care and Youth Development (NICHD ECCRN, 2000, 2002a, 2002b, 2005, 2006; NICHD ECCRN & Duncan, 2003), and two studies at the National Center for Early Development and Learning (Clifford et al., 2005; Pianta et al., 2005). Observation is being used to assess early childhood classroom quality in less standardized yet more evaluative ways through state policies; 38 states require site visits to monitor the quality of state-funded pre-kindergarten classrooms (Barnett et al., 2008) and observational measures are often involved as part of their Quality Rating and Improvement Systems (National Association of Child Care Resource and Referral Agencies, 2009; National Child Care Information and Technical Assistance Center, 2009).

In kindergarten-12th grade settings, direct observation is a common tool for school administrators and coaches to apply when evaluating and providing professional development for teachers (Dickinson & Caswell, 2007; Pianta & Allen, 2008). Also, the Bill & Melinda Gates foundation recently designated $22 million in grants for research on teacher effectiveness; a portion of this funding will be used to observe and videotape teachers (Medina, 2009; Robelen, 2008, 2009).

As direct observation grows in its use for evaluating and improving quality across a variety of educational settings, there is even potential for scores from direct observations to be used in high-stakes contexts. For example, teachers or programs may be required to meet a certain threshold of quality to retain funding, as is already the case through QRS in some states (National Child Care Information and Technical Assistance

Center, 2007). In this context, having raters who make valid, reliable observations is crucial. However, when the scope of observational assessment involves more than a single classroom or school, but instead engages all schools in a state or across multiple states, training and mobilizing a workforce of raters that can assign reliable scores can be challenging and expensive.

Two significant concerns for coordinators of large-scale observational assessments include the process of training large numbers of staff in a timely and effective manner, and hiring staff who are capable of observing in objective ways. First, can you train staff from varying backgrounds to perceive classrooms in the same way? Initial training for new raters typically takes several days (Halle & Vick, 2007) and scheduling these trainings can be difficult. In North Carolina, there is a waiting list for observational assessments just because there are not enough raters trained to visit programs in a timely fashion (Zellman & Perlman, 2008). Training raters in large numbers over a short period of time is a significant undertaking. Second, who do you hire to rate classrooms? States so far have used a variety of approaches to staff raters for Quality Rating Systems (QRS), contracting them through university settings, state licensing agencies, or child care resource and referral agencies (Stoney, 2004). Often, raters are asked to provide quality improvement support for programs as well as evaluate them (Zaslow et al., 2009; Zellman & Perlman, 2008), so finding people capable of serving dual roles can be a concern given the opportunity for rater bias. There are many other possible sources of rater bias as well (Johnson, Penny, & Gordon, 2008; Merrell, 1999) and it is important to identify those that pertain to observational assessment. Very little is known about the characteristics of raters that predict whether they can objectively observe classrooms and assign reliable scores on assessments.

The current project proposes to examine these questions by capitalizing on data collected through the Office of Head Start (OHS) in 2008-2009. The Improving Head Start for School Readiness Act of 2007 required OHS to include a valid and reliable observational tool for assessing classroom quality in grantee monitoring (U.S. Department of Health and Human Services et al., 2008). What followed was a nationwide effort to train practitioners on a research-based observational assessment, an effort which provided a special opportunity to study raters. Documentation of scores from calibration assessments following initial training makes it possible to assess the success of a scaled-up approach to training raters. A portion of the trainees also completed a brief survey of demographic characteristics, job responsibilities, and beliefs about teaching which provides a window on who was most successful in calibrating.

Drawing on this unique set of data, two research questions are addressed. First, the degree to which observers were able to calibrate to pre-determined ratings following an initial training session is described. Second, we explore the relationship between the degree of calibration and characteristics of observers or training sessions. We start by discussing the development of observation systems and some challenges to using them at-scale.

**Development of Observational Systems**

Tools for direct observation have been available for many years now and new ones continue to be developed today. In the late 1960s, a fifteen-volume anthology called *Mirrors for Behavior* was written to include information on 80 observation systems available at the time (Simon & Boyer, 1969). Today, many observation systems are available to school administrators to support their teachers' professional development that advertise brevity, informality, simplicity of use, or adaptability (e.g. Downey, English, Steffy, Poston Jr, & Frase, 2004; Montgomery, 2002; Zepeda, 2008). In one example, 40 "easy-to-use tools" are available in the form of a workbook, with blank observation sheets available in the back for notes and diagrams (Zepeda, 2008). Yet relatively few of these observation systems are supported by rigorous research or include information on reliability and validity (Hintze, 2005); some systems rely primarily on case studies for evidence (Colvin et al., 2009; Montgomery, 2002).

If ratings or notes from observations are to be compared across teachers within schools, or across teachers from different schools, a standardized protocol for observing and for interpreting the data is important, particularly when observations occur at a large scale. In the case of using observation for professional development, this contributes in part to establishing a common language for teachers and administrators (Pianta & Hamre, 2009). Even more important than establishing a common language, however, is establishing that data collected as part of an observation accurately reflects the behavior of interest and minimizes observer bias.

There are standardized observation systems available with demonstrated reliability and validity that are appropriate for large-scale assessment of teachers and classrooms. A number of them have been developed for use in early childhood education. For example, the Arnett Caregiver Interaction Scale (Arnett, 1989) has been used in large studies such as the Head Start Family and Child Experiences Survey (Zill et al., 2001; Zill, Sorongon, Kim, Clark, & Woolverton, 2006) and the Preschool

Curriculum Evaluation Research Initiative (Preschool Curriculum Evaluation Research Consortium, 2008) to rate teacher responsiveness, tone, and discipline style. The Early Childhood Environment Rating Scale (ECERS; Harms & Clifford, 1980) and its later revision (ECERS-R; Harms et al., 1998) have been widely used to assess child care quality in terms of whether a classroom has appropriate routines, activities, and materials in place for children, provisions for parents and staff, and whether teaching staff interact with children in developmentally appropriate ways. The Early Language and Literacy Classroom Observation (ELLCO) Toolkit (M. Smith et al., 2002) is another observation instrument that has been widely used to collect information about the materials available and a teacher's approach to facilitating children's language and literacy skills.

Reliability and validity of observational tools have been assessed in a variety of ways. The reliability of observation data is typically assessed by inter-observer agreement, the degree of consistency in scores across multiple independent observers, and by intra-observer reliability, an estimate of the consistency in a single observer's scores across multiple observation occasions (Hintze, 2005). There are many ways to calculate each of these two types of reliability. For example, inter-observer agreement can be estimated as the percentage of the total number of indicators for which multiple observers agree, or even a correlation between the two sets of scores. Alternatively, a kappa coefficient provides an estimate of inter-observer agreement that is corrected for the number of expected agreements on the basis of chance, and is considered a more conservative estimate of reliability. A variety of strategies for assessing intra-observer reliability have been adapted from techniques traditionally used to assess individual-level measures, including test-retest reliability and Cronbach alpha.

For example, the Classroom Assessment Scoring System (CLASS; Pianta, La Paro et al., 2008a) has been used to observe teacher-child interactions in more than 4,000 classrooms (Hamre et al., 2008). Inter-observer agreement is established first in a calibration test following initial training (observers must assign scores within one point of master codes) and on an ongoing basis by comparing ratings of the same session from multiple observers (observers must assign scores within one point of each other). In past data, inter-rater agreement following initial training is reported at 87% (Pianta, La Paro et al., 2008a, p. 95). For 33 videotapes rated as part of the MyTeachingPartner study, scores for two observers were within one point of each other for 78.8 – 96.9 percent of the scores, depending on the CLASS dimension. Concurrent and predictive validity have also been established for the CLASS. Observed CLASS scores in data from the National Center for Early Development and Learning's (NCEDL) Multi-State Pre-Kindergarten Study correlate with observed scores on the ECERS-R, with higher correlations ranging from .45 to .63 for the ECERS-R interactions factor and moderate correlations ranging from .33 to .36 for the ECERS-R furnishings and materials factor (Howes et al., 2008; Pianta et al., 2005). Also, higher CLASS scores are associated with children's academic, social, and behavior outcomes (Burchinal et al., 2009; Curby, LoCasale-Crouch et al., 2009; Hamre & Pianta, 2005; Mashburn et al., 2008).

**Challenges to Coordinating Observational Assessment**

Standardized observation systems have generally been developed in the context of large research studies like the Early Childhood Longitudinal Study – Kindergarten cohort (ECLS-K), the National Institute of Child Health and Human Development (NICHD) Study of Early Child Care, and the NCEDL Multi-State and SWEEP studies, and now, several observation instruments are commercially available. School administrators and state staff can purchase the trainings, manuals, and score-sheets they need to incorporate observation into quality evaluation and improvement systems. However, simply purchasing an observational system with demonstrated reliability and validity does not ensure that data collected by others using the tool will also be reliable and valid. Coordinators of observational assessments are given little guidance when charged with establishing a staff and a protocol capable of collecting reliable and valid data. To start, they must decide how to train raters, and who should be trained.

**Training raters.** Training appears to improve the capacity of observers to assign reliable ratings. In a meta-analyses involving 79 generalizability studies with information on raters, Hoyt and Kerns (1999) found that the magnitude of variance in observer ratings that is accounted for by observer bias can be reduced through moderate (5 to 24 hours) to high (25 or more hours) levels of training.

Typically, training raters involves reviewing the scoring rubrics, establishing benchmarks through examples, opportunities to practice scoring followed by discussion and feedback, and an assessment of calibration to scores assigned by expert raters (Johnson et al., 2008). Often, training continues until a rater passes the calibration assessment at a pre-determined criterion, such as 60-80% agreement with expert scores (Moon & Hughes, 2005; Penny, 2003).

Still, some observation systems may simply be more difficult than others for observers to learn to assign reliable scores (Merrell, 1999). For example, items that are part of an observation system may be defined too broadly, making inter-rater agreement difficult and/or decreasing the meaningfulness of the score, or too narrowly, requiring many items and becoming cumbersome for observers.

The level of training required to establish acceptable inter-rater reliability on observational measures varies depending on characteristics of the observation and of the observer, and can require intensive resources. Halle and Vick (2007) report training requirements in their compilation of early childhood education quality measures. Many of the measures listed require 2 to 3 days of training, followed by a reliability test, for observers to become certified. A quick scan of the compendium reveals at least 7 of the 35 measures listed fall into that category, including the CLASS, the Early Childhood Classroom Observation Measure (ECCOM; Stipek & Byler, 2004), and the Quality of Early Childhood Care Settings: Caregiver Rating Scale (QUEST; Goodson, Layzer, & Layzer, 2005). Other measures take much less or much more time for training. For the Child Care HOME Inventories (CC-HOME; Bradley, Caldwell, & Corwyn, 2003), training can take as little as ½ day, and the authors write that "it is not generally necessary to have such intensive training in order to achieve reliability on the CC-HOME" (Bradley et al., 2003, p. 300). Leff and Lakin (2005) report training requirements on a few measures that vary just as much, from 3 hours to 20 hours per week over 12 weeks.

Clearly, initial training on a standardized observation system can be a significant time investment, but it is a significant financial investment as well. The cost of training is quite variable, as expenses for manuals and trainers can range from nothing at all to $2500 to train 10 observers, excluding the cost of materials and travel expenses (Halle & Vick, 2007). This cost is in addition to the ongoing costs of following through on the system, factoring in the frequency of observations, the number of classrooms/programs sampled, the amount of time spent in each classroom, and ongoing reliability checks and support (National Child Care Information and Technical Assistance Center, 2007; Stoney, 2004; Zellman & Perlman, 2008).

Even if the demand of time and money for initial training is manageable, scalability can still be a challenge. For some observation systems, only the developers are able to train observers to score reliably, making trainer availability a potential barrier. For example, reliability training for ECERS-R must be specially arranged with the authors (e.g. Frank Porter Graham Child Development Institute, 2009). In other cases, interested parties can select a “train-the-trainer” approach and become qualified to teach the observation system to others (e.g. Teachstone, 2009). This approach can be used to facilitate training large numbers of raters at a faster pace, but leaves the authors of the observation system with less control over the quality of the training process.

**Who can learn to use observational assessment?** There is little data to predict who can use an observational system to objectively evaluate teacher quality following initial training, though some work has been done to examine qualifications for raters of other types of performance assessments. Ideally, raters possess an understanding of both the domain being assessed and the people involved (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1999).

There are several reasons why observers may have difficulty learning to assign reliable scores. Sources of rater bias have been explored as they apply to scoring writing assessments, and some of them may apply to assigning observed scores as well. For example, one study looked at the influence of writing style on raters' scores of content mastery (Schafer, Gagné, & Lissitz, 2005). While raters in that study were able to distinguish writing style from content mastery, one can imagine a parallel in classroom observation where raters have difficulty distinguishing a teacher's personality or interaction style from classroom practices. Rater bias may be introduced if his or her beliefs differ from standards set by the observation scoring rubric or the values evident in the classroom being observed (Johnson et al., 2008).

Merrell (1999) describes two potential sources of observer bias. First, observers' scores may be affected if they or their supervisors have a vested interest in the outcome of the observations. Second, observers will be more likely to observe a behavior they expect to happen. These potential threats to validity are difficult to discern in practice, but they speak to the importance of careful training for observers.

There are many occasions when observers could have a vested interest in the outcome of observations, and intentionally or unintentionally affect the assigned scores. Observers who also supervise or support the teachers they are rating may be influenced by their relationships with the teachers or additional information about the classroom not pertinent to the observation session. For example, this could be the case if program directors are in charge of observations. In some states, including North Carolina and Ohio, consultants teach directors about observational assessments so they understand the system and know how their centers can improve (Frank Porter Graham Child Development Institute, 2003; Zellman & Perlman, 2008). Also, raters in many states serve dual functions when they visit classrooms, as both evaluators and quality support staff (Zaslow et al., 2009; Zellman & Perlman, 2008). This typically happens when states that are struggling to finance a QRS combine job responsibilities in efforts to save resources (National Association of Child Care Resource and Referral Agencies, 2009; Zellman & Perlman, 2008). Alternatively, raters who are only visiting programs to evaluate may still have difficulty maintaining objectivity when they feel responsible for putting teachers at ease with the process. In Minnesota, raters also see themselves as "ambassadors" of the QRS, and try to build rapport with program staff when they visit (Tout et al., 2008, p. 26).

In school contexts, there are many reasons why an observer may expect to see or not see a specific behavior. Observers who have a great amount of past experience teaching or supervising classrooms may unknowingly be influenced by their own ideas regarding instruction or classroom management. In QRS, raters can be drawn from universities or child care resource agencies, as has been the case in North Carolina, Oklahoma, Tennessee, and Pennsylvania, or from state agency staff (Stoney, 2004). These people likely vary in their levels of experience and training in early childhood education. The degree to which variation in education and work experience influences observation skills is yet unexplored.

### When Observation Occurs At-Scale

The challenges described above of how to train raters and who to train are magnified when observational assessment is used in large-scale contexts. Implementation becomes a major challenge, as there are typically constraints placed on time and resources such that project coordinators want raters to be trained quickly and cheaply, but also effectively. When raters are trained for a single research project in an academic institution, having 20% of raters fail an initial calibration assessment translates into additional support required for a handful of raters, which can usually be dealt with by a single coordinator over the course of a few days. But if 20% of raters in a large-scale effort fail an initial calibration assessment, follow-up support could require much more significant investments of time and money, potentially involving many coordinators and weeks of effort. Knowing how to train large numbers of raters quickly and effectively is really important for at-scale efforts to operate smoothly.

Effective trainings are also crucial in large-scale contexts because raters who are not fully calibrated to an observational tool can introduce error into the data they collect. Measurement error of any kind, including error related to rater bias, can reduce statistical power for evaluating the effects of interventions and introduce bias to estimates of the relationships between observed scores and outcomes of interest (Raudenbush & Sadoff, 2008). This is problematic for researchers who need the data collected by raters in large-scale projects to confirm and further explore questions regarding teacher quality. This is also problematic for teachers and administrators who are subject to policy and retention decisions based on scores assigned by these raters. For all of the effort required to coordinate observational assessment in large-scale contexts, and for the impact of decisions resulting from these efforts, knowing how to create a workforce of calibrated raters and who to select for training are foundational first steps.

### The Current Study

Given the widespread use of observation in school contexts by administrators, professional development mentors, teachers, and researchers, this project takes a closer look at the characteristics of people who are able to assign reliable scores on an observational measure of teacher-child interactions, the Classroom Assessment Scoring

System (CLASS; Pianta, La Paro et al., 2008a) and the success of the scaled-up approach for training them.

Per the Improving Head Start for School Readiness Act of 2007, the Office of Head Start (OHS) was required to include a valid and reliable observational tool for assessing classroom quality in grantee monitoring reviews (U.S. Department of Health and Human Services et al., 2008). OHS chose to pilot using the CLASS for this purpose in 2008-2009. Accordingly, 2,117 Head Start administrators across the United States were invited to be trained to assign reliable CLASS scores and then to use the CLASS for observation and professional development in their own programs if they so chose. It should be noted that the staff who participated were not given explicit directions on what they would be doing with the CLASS tool once they completed the training sessions. Instead, CLASS trainings were presented as opportunities to build staff capacity to use the CLASS to assess and improve classroom quality in their programs, and use of the CLASS as an observational instrument or professional development tool by grantees was strictly voluntary. Still, this was one of few large-scale efforts to train practitioners on a research-based observation tool and was also an opportunity to collect information on the effectiveness of the training for this population. As a result, we can examine whether participants in the scaled-up training approach learned to assign reliable scores, and whether characteristics of these participants predicted their level of calibration.

Assigned scores were collected for all participants who completed the calibration assessment following the training. From this data, we will describe the degree of calibration for large numbers of raters trained in an at-scale framework, and trends in scoring among naïve observers. Some trainees also completed a short survey in which they reported on demographic characteristics (e.g. education background) and job responsibilities. This data will be used in conjunction with the calibration data to identify rater characteristics that are associated with greater calibration to observational tool.

## Method

### Participants

All Head Start grantees and delegate agencies were invited to send staff to participate in regionally-based, 3-day CLASS trainings. The number of staff eligible to participate per program was based on the number of children served by that program. One staff member was eligible for programs serving less than 500 children. Two staff members were eligible for programs serving between 500 and 1,000 children. Head Start programs serving 1,001-2000 children were allowed to send three staff members, and any programs serving more than 2,000 children were allowed to send four staff members.

Ultimately, 2,117 Head Start staff members participated in the CLASS trainings. The trainings were designed for staff whose job responsibilities included the supervision and/or professional development of teaching staff. The vast majority of these participants attempted the calibration assessment; only 1% of participants chose not to or were unable to complete it. Of 2,093 participants who did complete the calibration assessment, 704 also elected to complete a brief survey at the beginning of the training session in which they reported on demographic characteristics, job responsibilities, and beliefs about teaching. Those who completed the survey were entered into a raffle for a $10 gift certificate. Among those who completed the survey, 13% have an associate’s degree or less, 48% have a bachelor’s degree, and 37% have a master’s degree or higher. They reported an average of 9 years of experience supervising and/or mentoring teachers and

43% have been in their current position with Head Start for 1-5 years. Staff were predominantly female (95%) and were 47 years old on average (*SD*=10.07).

**Procedure**

CLASS trainings were led by 25 trainers who were also Head Start Training and Technical Assistance (T/TA) Specialists responsible for working directly with grantees to support their meeting monitoring and performance standards. All TA Specialists attended a 5-day Train-the-trainer workshop provided by the University of Virginia (UVA). The workshop involved training the specialists to reliability on the CLASS and then providing additional information and support required to become a certified CLASS trainer. A research scientist from UVA was available to trainers for ongoing support, including providing CLASS-related information to prepare for trainings, partnering with the specialists during initial trainings to answer questions and provide feedback, and stepping in as a substitute trainer when necessary.

There were 121 CLASS trainings that took place between November 2008 and August 2009. Up to 20 participants were allowed to attend each. Materials were made available to participants beforehand, including a CLASS manual and 1-month access to the CLASS website (www.classobservation.com). The CLASS website offers basic information about the CLASS and registered viewers are also able to view hundreds of short videos of interactions among teachers and students. Each video segment is included to illustrate a specific component of high quality teacher-child interactions according to the CLASS and is accompanied by a brief explanation.

The trainings were each three days long. In the first two days, the CLASS structure and coding protocol was introduced and trainees practiced coding five 20-

minute video segments of real preschool classrooms. There was time for trainees to ask questions and engage in discussion to further develop their understanding of the CLASS and the coding process. The calibration assessment took place on the third day. For this assessment, trainees watched three 20-minute video segments and spent 20 minutes coding each one. Following the calibration assessment on the third day, trainers provided information on CLASS-related professional development and research and then opened the floor to group discussion of these ideas and suggestions for using the CLASS to reach participants' professional development goals.

Feedback from training sessions that occurred in November and December 2008 was used to adjust the trainings starting in January 2009. Changes to the training format included allowing more time for discussion by reducing the number of between-training activities, adding a "practice calibration" segment at the end of the second day for which they received individualized feedback from trainers before starting the CLASS calibration assessment, providing trainees with calibration scores before starting the professional development portion of the trainings, and reducing the number of video clips viewed during the professional development portion of the training.

For trainees who were unsuccessful in their first attempt to pass the calibration assessment, opportunities for two additional attempts were available through the University of Virginia's (UVA) Center for Advanced Study of Teaching and Learning. Interested trainees were provided contact information for a UVA staff member who gave them online access to additional calibration segments for a $20 fee. Second and third attempts at the calibration assessment also involved viewing three 20-minute segments and submitting codes.

## Measures

**CLASS.** The Classroom Assessment Scoring System (CLASS; Pianta, La Paro et al., 2008a) is an observational assessment of teacher-child interactions that has been validated through use in more than 4,000 classrooms (Burchinal et al., 2008; Hamre & Pianta, 2005; Hamre et al., 2008; Mashburn et al., 2008). Observers assign global ratings on a scale of 1-7 to each of ten dimensions of teacher-child interactions. Observers are expected to determine ratings from specific, behavioral examples that reflect both the quantity and quality of teacher-child interactions relevant to each dimension, with guidance from the text of the CLASS manual.

On the CLASS scale, scores of 1-2 are considered low-range quality of teacher-child interactions, 3-5 are mid-range, and 6-7 are considered high-range quality. The ten dimensions factor into three domains of interactions: Emotional Support, Instructional Support, and Classroom Organization. Emotional Support represents the average of scores from the dimensions of Positive Climate (warmth and respect among teacher and students), Negative Climate (reversed, teacher and peer negativity), Teacher Sensitivity (teacher awareness and responsiveness), and Regard for Student Perspectives (emphasis on student interests and autonomy). Instructional Support is calculated from the dimension scores of Concept Development (focus on higher-order thinking skills and understanding), Quality of Feedback (feedback expands learning and encourages participation), and Language Modeling (teacher models and facilitates language). Finally, Classroom Organization is calculated from the dimensions of Behavior Management (effective methods to prevent and redirect misbehavior), Productivity (effective management of instructional time and opportunities to learn), and Instructional Learning Formats (facilitation of student engagement and learning).

**Rater calibration.** For the current study, we were interested in trainees' calibration scores on an assessment following the CLASS observation training. Trainee calibration scores were calculated from their accuracy in rating dimensions for three 20-minute segments. To pass the calibration assessment, trainees were required to meet the following criteria: 1) 80% of all codes had to be within 1 of master coded scores, 2) 1 out of the 3 possible scores were correct (within 1 of the master coded scores) in every CLASS dimension. Of particular interest was the percentage of trainee-assigned CLASS scores that fell within 1 point of master-coded scores across all three calibration segments. This is the overall adjacent calibration score and is the measure of calibration most typically employed by projects using CLASS. It is calculated by dividing the number of trainee-assigned scores that are within 1 of the master-coded scores by 30, the total number of scores to assign.

However, there are many alternatives to this standard measure of calibration. First, adjacent reliability can be calculated for each dimension independently, as well as across all ten dimensions. Second, distance scores can be calculated across all ten dimensions and for each dimension independently. Distance scores represent the gap between a trainee-assigned code and the CLASS master-code and are averaged across all three reliability segments.

Distance scores were calculated in two ways. First, valence distance scores were calculated by subtracting the master code from the trainee rating. Thus, positive distance scores indicate that trainees' codes were higher than master codes (i.e. trainee assigned a

5 when the master code was a 4) and negative distance scores indicate that trainees assigned codes that were lower than master codes (i.e. trainee assigned a 3 when the master code was a 4). Second, distance scores were calculated as the average absolute distance from the master code across the three segments. Higher absolute distances indicate that trainees assigned codes that were further from the master codes, irrespective of direction.

**Trainee survey.** Starting in January 2009, CLASS training participants who agreed to participate completed a brief survey. This survey contained questions regarding demographic information, such as race/ethnicity and educational background. It also included questions on job responsibilities, such as the number of years of supervisory/mentoring experience, the amount of time spent observing teachers per week, and use of classroom observation systems in the past year. Finally, the survey included questions about a rater's beliefs regarding children and teaching practices.

**Additional information about training sessions.** At the trainings, data was also collected on the number of participants present and the date of the training.

**Analytic Plan.**

**Describing calibration.** The calibration assessment was completed by 2,093 of the 2,117 Head Start staff who participated in CLASS trainings. Three calibration metrics were calculated from rater-assigned scores for each of the master-coded video observations in the assessment. Descriptive statistics are presented for the adjacent calibration, valence distance scores, and absolute distance scores across all ten dimensions as well as for each dimension individually. This analysis describes the extent of rater calibration in a large-scale training effort, and valence scores illustrate trends in scoring among naïve observers.

**Prediction models.** Multilevel modeling was used to explore the characteristics of raters and training sessions that were associated with rater calibration. One of the three calibration metrics was selected for this purpose and the absolute distance calibration scores were preferred to the other metrics for several reasons. Absolute distance scores were selected over the adjacent calibration outcomes due to greater variation present in the absolute distance scores. Also, finding from models with the absolute distance scores are more easily generalized to other observation systems. Instead of the somewhat arbitrary 80% cutoff score for passing adjacent calibration, absolute distances reflect greater or lesser calibration to an exact agreement standard. The valence scores were not used in prediction models because interpretation of the coefficients is difficult. While valence scores for the full sample speak to trends in scoring, it is impossible to distinguish individual rater under- or over-scoring relative to the master-code from broader sample-wide trends without sub-sampling raters who did not meet preset criteria for passing.

The data for this study were hierarchically nested in three levels, with raters nested in training sessions, nested in trainers. Level-1 (rater) predictors were only available for the 704 raters who completed the survey and will only be included in models with that subsample. In this subsample, 704 raters were nested within just 84 of the 121 training sessions, and 24 of 25 trainers. Level-2 predictors will be tested with the survey subsample as well as with the full sample of 2,093 raters who completed the calibration assessment. All predictors were entered simultaneously, as we did not have preconceived ideas that certain predictors were more important than other ones. Also, all predictors were entered as fixed effects.

I had originally planned to calculate estimates of the association between individual trainee calibration and predictors at each of the three levels through Hierarchical Linear Modeling (HLM; Raudenbush & Bryk, 2002) using SAS Proc Mixed (Singer, 1998) for three-level models. However, preliminary analyses indicated that two-level models were more appropriate. In the survey dataset of 704 raters, there was only significant ($p$<.05) level-3 variance for 1 of 11 outcomes, the Language Modeling absolute distance score. In the full dataset, there was significant level-3 variance for Language Modeling, Concept Development, and Quality of Feedback. To simplify comparison across outcomes, I used a two-level model (raters nested in sessions) for all outcomes with dummy codes representing trainers at level-2 to control for any trainer effects.

***Level-1 predictors.*** The vast majority of predictors were available at the rater level. Race/ethnicity were entered as dummy codes for Black (18%), Latino (12%), and other (15%), relative to White (66%). Two predictors describe trainee's education experiences: whether a rater has a master's degree or more (37%), and whether the rater has had three or more courses that covered early childhood development/child development (71%).

Additional level-1 predictors describe trainees' job responsibilities. These include the number of years of supervisory/mentoring experience (*M*=9.26, *SD*=6.88). There is also data on trainees' experience with observational assessment, including the amount of time per week spent observing teachers (38% answered 6 hours or more), and whether they have used classroom observation systems in the past year, such as the such as the Early Childhood Rating Scale - Revised (41% yes; ECERS-R; Harms et al., 1998).

Raters also completed 16 items from the Modernity Scale(Schaefer & Edgerton, 1985) which examines teachers' beliefs about children. The items help differentiate teachers with "traditional" or relatively adult-centered perspectives on interactions with children and teachers who have more "modern or progressive" child-centered perspectives. Raters with adult-centered perspectives agreed with statements such as "Children must be carefully trained early in life or their natural impulses make them unmanageable." Teachers with child-centered beliefs agreed with statements such as "Children should be allowed to disagree with their parents if they feel their own ideas are better." On a scale from 1 to 5, with 5 being more adult-centered beliefs, the mean score across raters was 2.00 (*SD* = 0.55). Cronbach's alpha was 0.75 in this sample.

Finally, raters completed 11 items that examined their beliefs about intentional instruction (Downer & Hamre, 2010). Raters with more intentional teaching beliefs agreed with items such as "Young children learn best when teachers are actively involved with their play." Cronbach's alpha was 0.58 for this scale. However, factor analysis indicates that a subset of six items can be used (LoCasale-Crouch, Downer, & Hamre, 2010). In a previous sample Cronbach's alpha was equal to 0.75 for these six items; in the current sample, Cronbach's alpha is equal to 0.72 for this subset. On a scale from 1 to 5, with 5 being more intentional teaching beliefs, the mean score across raters was 3.74 (*SD* = 0.50) for the subset of six items.

***Level-2 predictors.*** Predictors of interest at the training session level included the date of the training, formatted as days since November 1, 2008, the month of the first training sessions. The challenge of the group as a whole is captured in predictors that include the number of participants present, and in the survey data, aggregated trainee characteristics that include their beliefs about children and intentional teaching. Finally, dummy codes representing individual trainers were included to control for any trainer effects. The trainer with the highest adjacent calibration rate from their own training experience was selected as the reference group; the average adjacent calibration on first attempt was 77.44 for trainers (*SD*=8.24), and ranged from 60-92%.

***Final models.*** For the survey data, the absolute calibration score for rater i in training session j includes the intercept or overall calibration score, plus the contributions of race/ethnicity, master's degree, coursework in early childhood development, years of supervisory experience, hours per week spent observing, prior use of the ECERS-R, beliefs about children, and beliefs about intentional teaching. The intercept is further defined at Level-2 by controls for trainer, the number of days since trainings began, the grand-mean-centered number of raters per training session, beliefs about children and teaching aggregated to the session-level, plus error for session j. Intercept differences were allowed to vary across sessions, but the effects of $\beta_{1\text{-}10}$ were fixed at the session-level.

$Y_{ij} = \beta_{0j} + \beta_{1j}$ (Black) + $\beta_{2j}$ (Latino) + $\beta_{3j}$ (Other race) + $\beta_{4j}$ (MA or higher level of education) + $\beta_{5j}$ (three or more courses in early childhood development) + $\beta_{6j}$ (years of supervisory experience) + $\beta_{7j}$ (six or more hours per week spent observing) + $\beta_{8j}$ (used ECERS-R in past year) + $\beta_{9j}$ (beliefs about children) + $\beta_{10j}$ (beliefs about intentional teaching) + $r_{ij}$

$\beta_{0j} = \gamma_{00} + \gamma_{01\text{-}23}$(dummy-codes for trainer) + $\gamma_{24}$(days since Nov. 1, 2008) + $\gamma_{25}$(number of raters) + $\gamma_{26}$(beliefs about children) + $\gamma_{27}$(beliefs about intentional teaching) + $u_j$

$\beta_{1\text{-}10j} = \gamma_{1\text{-}10}$

Level-1 predictors were excluded from models using the full dataset, for the following equations:

$Y_{ij} = \beta_{0j} + r_{ij}$

$\beta_{0j} = \gamma_{00} + \gamma_{01\text{-}23}$(dummy-codes for trainer) + $\gamma_{24}$(days since Nov. 1, 2008) + $\gamma_{25}$(number of raters) + $u_j$

**Missing data.** There were low levels of missing data in both the full data reflecting 2,093 raters and the survey data reflecting 704 raters. Dimension scores from the calibration assessment were missing for less than 1% of trainees. Individual survey items were missing at rates of 3% or less. Because levels of missing data were low, I considered dealing with them through listwise deletion. Deleting all cases with missing data on any variable of interest would only diminish the full dataset by 4%, from 2093 trainees to 2002 trainees. However, the same process in the survey dataset would decrease the number of cases from 704 to 445, a loss of 37%.

Multiple imputation is a more appropriate approach to handling missing data in the survey dataset, and will be conducted for the full dataset as well to be consistent across analyses. Ten imputations were calculated and final results were summarized across imputations. Imputation was conducted separately for the full and survey datasets, and the survey variables were not included when imputing for the full dataset. Also, I did not re-calculate overall adjacent calibration scores for each imputation. This variable is non-missing across all cases in the original data and is likely to reflect the true overall adjacent calibration score even when dimension-level scores are missing due to administrative error in tracking the data.

## Results

### Describing Calibration

Descriptive statistics for the three calibration metrics are presented in Tables 1 through 3. The majority of raters passed the calibration assessment according to preset criteria on their first attempt; 71% of raters assigned at least 80% of codes within-1 of master-codes. Adjacent calibration rates varied by dimension (see Table 1), with the poorest performance occurring for the three dimensions in the Instructional Support domain, Concept Development, Quality of Feedback, and Language Modeling. Raters who completed surveys performed slightly better on the overall adjacent calibration (Mean = 84.50) than raters who did not (Mean = 81.97). This difference was statistically significant ($t$ = -15.24, $p$<.0001).

Valence distance scores varied by dimension (see Table 2). There was a trend of assigning codes higher than the master-code for the three dimensions in the Instructional Support domain, Concept Development, Quality of Feedback, and Language Modeling, ranging from .71-1.05 points above master-codes. Raters tended to over-score Regard for Student Perspectives as well, by about half of a point. Under-scoring happened in few cases and not to the same degree.

The larger valences for the Instructional Support dimensions are reflected in the average absolute distance scores as well. Absolute distances ranged from .79-1.07 points away from master codes for these dimensions, and were followed in magnitude by

Regard for Student Perspectives at .61. Still, raters were not wildly off from master-codes, 95% of raters assigned scores that averaged within-1 point of master-codes, including raters who did not pass according to the adjacent calibration criteria.

### Predicting Calibration

Two-level models were used to explore whether absolute distance scores were predicted by characteristics of raters and the training sessions they attended. When effect estimates are negative, they indicate shorter distances between rater codes and master-codes. When estimates are positive, they indicate that raters were further off.

***Full sample.*** Level-2 predictors were included in models using the full sample of 2,093 raters who completed the calibration assessment (see Tables 4a-d). There were trainer effects in some cases and many trainer effects for Positive Climate. Above and beyond trainer effects, the date of the training appeared to matter. Raters assigned scores that were closer to master-codes in training sessions that occurred later in the year, significantly so for the dimensions of Negative Climate, Behavior Management, and Concept Development. There were trends toward smaller distances for Teacher Sensitivity and Behavior Management when more raters than average attended training sessions, but these effects did not reach statistical significance.

Level-2 predictors accounted for substantial session-level variance in many cases. Predictors accounted for 39, 34, and 38% of the level-2 variance in the unconditional model for the outcomes of Concept Development, Quality of Feedback, and Language Modeling respectively. They also accounted for 62% of the level-2 variance for Teacher Sensitivity. On the contrary, they accounted for none of the level-2 variance for Productivity.

***Survey sample.*** Level-1 predictors were added to models using only the sample of 704 raters nested within 84 training sessions who completed surveys as well as calibration assessments (see Tables 5a-k). Broadly speaking, raters' beliefs about children and teaching appeared most important in accounting for variation between raters, and raters' level of education or job responsibilities did not matter very much.

More specifically, raters who held more adult-centered beliefs were further off from master-codes in Concept Development and Language Modeling. Raters with more intentional teaching beliefs assigned scores that were closer to master-codes overall, and for a number of dimensions: Regard for Student Perspectives, Concept Development, Quality of Feedback, and Language Modeling. There were other significant fixed effects for race/ethnicity. Raters who are Black, Latino, or another race tended to be more off from master-codes than raters who are White for dimensions in the Instructional Support domain.

Data on education or job responsibilities did not consistently predict absolute calibration scores. There were a few exceptions. Namely, raters who possessed a graduate degree were better at assigning scores for Concept Development. Raters who had used the ECERS-R in the past year were better at assigning scores for Negative Climate. Finally, raters who typically spend 6 or more hours observing classrooms each week were worse at scoring Language Modeling.

Similar to models with the full sample, there were few significant level-2 predictors in the models with the survey data. There were some effects for trainer, and raters assigned scores closer to master-codes at training sessions that occurred later in the year. Most interestingly, rater beliefs about children was aggregated and added as a predictor at level-2 in these models, and significantly predicts absolute distance for the overall and Regard for Student Perspectives scores.  When the group of raters attending a session tends to hold more adult-centered beliefs, raters are further off from master-codes.

The addition of level-1 and level-2 predictors in these models accounted for substantial rater-level variance for some outcomes, and barely any variance for others. For example, predictors accounted for about half of the level-1 variance for Instructional Learning Formats, Regard for Student Perspectives and Language Modeling.  They also accounted for about 40% of level-1 variance for Concept Development and Quality of Feedback.  In other cases, the predictors accounted for very little, if any variance at all.

## Discussion

Observational assessment can be used to identify teacher practices that are associated with children's academic, social, and behavioral outcomes (Burchinal et al., 2008; Burchinal et al., 2009; Curby, LoCasale-Crouch et al., 2009; Howes et al., 2008; Mashburn et al., 2008; Rimm-Kaufman et al., 2009) and to provide material for feedback and professional development (Dickinson & Caswell, 2007; Pianta & Allen, 2008; Pianta & Hamre, 2009).  For these reasons, observation tools are now being incorporated into large-scale efforts to measure and improve classroom quality, as is the case in many Quality Rating Systems evaluating the quality of early childhood education (Tout et al., 2009). When used in these ways, having standardized observation tools is really important, that evaluators, raters, teachers, and administrators can have a common metric for assessing and improving quality.  However, training people to view classrooms in standardized ways can be challenging and expensive, both in the sheer logistics of implementing trainings and in identifying people who are right for the job. When implementing large-scale projects, evaluators want rater training to take just the typical few days, with limited follow-up, but lack information on feasibility. Evaluators often want to use their own staff for leading trainings, because developers of observation tools are typically university-based and not equipped to train hundreds or thousands of raters, but staff must still be well-versed in the tool and it is unclear whether this is possible. Finally, evaluators want to be able to hire people who can calibrate to the tool right away without lots of expensive follow-up, but have few guidelines on identifying the best candidates from a pool that involves great diversity in experiences and beliefs about teaching.

Because most observation tools have been developed and used in smaller projects requiring fewer than 100 raters, opportunities to examine issues of rater calibration relevant to large-scale contexts have been limited. Data from the OHS efforts to train more than 2,000 raters on the CLASS allowed us to explore the extent of rater calibration when training to use an observation tool occurs at large-scale, as well as the characteristics of raters and training sessions that are associated with calibration. Overall, we saw that it is possible to train large numbers of raters to calibrate to an observation tool through 2-day training sessions led by the evaluator's own staff. Also, certain characteristics of raters were associated with this calibration that might inform rater selection for future at-scale efforts. The results and implications for planning and implementation of large-scale assessments are further discussed below.

Of Head Start staff trained on the CLASS tool, 71% passed the calibration assessment upon their first attempt. This finding tells us that it is possible to create a workforce of calibrated raters, but not all raters will pass a calibration assessment on the first try. The passing rate should be interpreted with caution, given the context that the purpose of these CLASS trainings was not to establish a monitoring system, but to instead familiarize grantees with the CLASS tool and create potential for its use in assessing and improving classroom quality in their programs. Actual use of the CLASS for these purposes was completely voluntary. Still, the rate does provide insight for planning at-scale observational work.

When planning large-scale protocols, coordinators should consider provisionally hiring more raters than they eventually need and only retaining those who pass the calibration assessment. Johnson and colleagues (2008) recommend bringing in 20-30% more raters for initial interviews than needed for a project, and the results here are in line with that estimate. Alternatively, coordinators of large-scale observational assessments could hire only as many raters as they actually need, but plan to provide extra support for a percentage of raters as they continue to attempt to calibrate to the tool. It is difficult to make recommendations regarding the exact percentage of raters for which evaluators should plan to over-hire or provide additional training because these numbers may depend on the context for the training. When raters in this study were trained, they were told that use of the CLASS in their own programs was voluntary. It is certainly possible that passing rates would be different if raters were being trained to do evaluation work in a high-stakes context.

The results from this study also give us confidence that the “train-the-trainer” approach to organizing trainings can be successful in a large-scale context. Head Start Specialists were trained by staff from the University of Virginia to lead CLASS trainings for other Head Start staff. Results from multilevel models indicated that the vast majority of variation in rater calibration scores occurred between raters, and not between trainers or even training sessions. When trainers are provided with specific instructions for training, any variation that does occur between training sessions is minimally reflected in raters' calibration scores.

It does appear that some components of an observation instrument can be more difficult for raters to calibrate to than other ones, just as some instruments may be more difficult to learn than others (Merrell, 1999). Rater-assigned scores were close to master-coded scores on average, but absolute distance scores were the largest for dimensions in the Instructional Support domain (Concept Development, Quality of Feedback, and Language Modeling), with Regard for Student Perspectives following behind them. In these cases, raters appear to be assigning scores that are higher than master codes, weighing certain examples of teacher-child interactions too heavily in assessing their quantity and quality. While this may be explained by greater variation present for these outcomes, it is also consistent with anecdotal reports that the Instructional Support domain is difficult to teach and to learn. One possibility is that the definitions for these dimensions are lacking in clarity (Merrell, 1999). Alternatively, it is possible that these dimensions of the CLASS reflect a way of thinking and talking about teaching that is slightly different from common understanding, and requires more of a shift in beliefs and knowledge for naive raters, and in this case raters who are also practitioners, to calibrate. When rater beliefs are misaligned with the theoretical foundation for an observational assessment, calibration could be problematic.

Indeed, raters' beliefs about children and teaching were the most consistently predictive of rater calibration, above and beyond rater education and experience. This was especially true for those dimensions where the absolute distance scores were largest. In this example, raters who believed intentional teaching practices are important were more closely calibrated with CLASS dimensions, including those dimensions in the Instructional Support domain and Regard for Student Perspectives. The CLASS focuses on interactions between teachers and children, and the specific practices that teachers use to facilitate children's activities and learning in classrooms. Raters whose own beliefs about the role of a teacher in a classroom were aligned with the CLASS approach were more calibrated in the initial assessment. Putting aside personal biases and opinions about teaching was less of an issue for these raters, as their beliefs were in line with those emphasized by the CLASS. These results suggest that evaluators should attend to the alignment of rater beliefs and the underpinning theories of an observational assessment in order to minimize rater bias.

Moreover, the average beliefs of the group of raters participating in a training session also influenced rater calibration. When the beliefs of a group of raters were more adult-centered on average, raters were less calibrated overall, and particularly for Regard for Student Perspectives. Similar to the findings for beliefs at the rater-level, this result indicates that alignment between the focus of the observation tool and the beliefs of the raters can be important. The dimension of Regard for Student Perspectives reflects the degree to which a teacher's interactions with children facilitate children's interests and autonomy in the classroom. Raters who hold more adult-centered beliefs about teaching may strongly disagree with the positioning of child-focused interactions as high quality.

The finding that adult-centered beliefs predicted less calibration at the session-level suggests that discussions become difficult when more raters within a session disagree with what is being presented. Perhaps when the majority of participants in a training session carry certain values, it is more difficult to train them to perceive teacher-child interactions in a way that differs from those values.

The findings that rater beliefs are important, particularly for certain dimensions, have implications both for how raters are trained and who should be trained to do observational assessments at-scale. If components of an observation tool are controversial, or somehow different from common rater belief and understanding, relatively more time should be spent on these components during training. Additional discussion can be used to fully reveal rater bias and remind raters to set contrary beliefs aside for the purpose of the assessment. In CLASS training sessions, this translates to allowing plenty of time for thorough discussion of the dimensions that appear most sensitive to rater bias due to beliefs about teaching and children, including Regard for Student Perspectives and dimensions in the Instructional Support domain. Also, research coordinators should pay attention to raters' beliefs when making initial hiring decisions. When hiring staff primarily for observation responsibilities from a pool of applicants with diverse backgrounds, it appears that rater beliefs are more relevant than standard qualifications such as level of education or experience. Rater beliefs that are poorly aligned with tenets of the observational system at hand should be considered as warnings of potential rater bias in assigned scores.

A few other predictors were significantly associated with absolute distance scores as well. Raters' race/ethnicity predicted absolute calibration across several outcomes.

Generally, raters were less accurate in assigning scores for dimensions in the Instructional Support domain if they were non-White. It is not clear why this may be the case above and beyond the other predictors in the models, but it is possible that these predictors reflect cultural differences in beliefs about children or teaching that are unmeasured in this sample. It is also possible that raters were reacting to the race/ethnicity of the teachers and children in the classrooms they were observing. The lead teacher was White in all three video segments of the calibration assessment. In two of the three videos, the children were mostly Latino with some Black children, and in the third, the children were mostly White with some Black children. Likewise, the lead teacher was White in three of the five training video segments, and children across the five videos were generally White or Hispanic, with Black children being represented to a lesser degree. The lack of match between the race/ethnicity of raters and that of the teachers and children being observed could have contributed to the lesser calibration of non-White raters across dimensions in the Instructional Support domain.

There were also a few outcomes for which time since the training sessions first started predicted rater calibration at the session-level. This was true in the full dataset, but did not hold up when models only included raters who completed surveys. It is not surprising that the effects disappeared in the survey data, given that surveys were only available to raters who participated in training sessions that took place during or after January 2009. The most significant changes to the training sessions occurred simultaneously with the introduction of the survey and we would not expect the date of a training session to predict calibration after this point. Ultimately, that time since trainings began was positively associated with calibration indicates that calibration improved with the expansion of opportunities for discussion and practice during training.

***Limitations.*** A primary limitation of this study is that characteristics of raters that predicted calibration could only be examined for raters who completed the survey, when we know that raters who selected to complete the survey calibrated to the CLASS to a greater degree than raters who did not. Also, raters who were trained in November and December were not given the option to take the survey at all. It is impossible to know the other ways that these groups of raters differed.

Second, because some trainers led only one training session each, it is difficult to know if session-level effects should be interpreted as such, or if they should instead be interpreted as trainer-level effects. Across 121 sessions and 25 trainers in the full sample of data, each trainer led a minimum of 1 session, a maximum of 15 sessions, and the median number of sessions was 4. Across 84 sessions and 24 trainers in the survey sample of data, each trainer led a minimum of 1 session, a maximum of 15 sessions, and the median number of sessions was 4.5. The total number of sessions a trainer led was not significantly correlated with any of the absolute distance outcomes in the prediction models, or with the percentage of raters who had at least 80% of scores within-1of master-codes, in either dataset. Still, some caution should be used in interpreting session-level effects given the possibility that session and trainer effects are confounded.

***Conclusion.*** This study has addressed issues relevant to rater calibration in large-scale contexts in a number of ways. First, we have learned that a scaled-up “train-the-trainer” approach to calibrating raters can work. The majority, but not all, of raters do calibrate after just two days of training. Second, certain characteristics of raters predict the degree of calibration after initial training. Rater beliefs are more consistently related to calibration than levels of education or experience.

Future implementers of large-scale observational assessments can be reassured that it is possible to train large numbers of raters to calibrate to an observational tool in a short period of time. When hiring raters, evaluators should pay special attention to rater beliefs and particularly the ways in which rater beliefs are or are not aligned with the observation tool at hand. Also, if specific components of an observation tool are controversial or offer a way of thinking that is less than common knowledge, evaluators should allow for sufficient time during trainings to expose and disperse rater bias associated with those components.

There are many opportunities for future research in this area. For one, while it is clear that the majority of variation in rater calibration is present at the rater-level, relatively little of the variation is accounted for by the present predictors for many of the outcomes. Additional sources of variation, unmeasured here, could include other beliefs and knowledge regarding early childhood education, or cognitive skills such as the ability to attend to relevant information and inhibit distraction. Whether or not a rater has teaching experience could also be important.

There are also questions about follow-up support for raters who fail initial calibration assessments. Project coordinators who struggle to decide how much follow-up to do with raters who fail initial calibration assessments would benefit from knowing the percentage of raters who succeed at later assessments after failing the first. What are the characteristics of raters who continue to fail calibration assessments after multiple opportunities for training and feedback? Are there some people who will never be successful at calibration, and what characterizes them?

Finally, it would be interesting to look more closely at what is occurring within trainings. Are rater concerns that stem from beliefs which are misaligned with an observation tool actually aired and alleviated through discussion? Is extended discussion successful in minimizing rater bias, or does it distract and make calibration more difficult for other raters who are present?

These questions become increasingly important as observational assessment becomes a popular tool for assessing and improving teacher quality. Standardized observation can be fundamental in establishing a language for assessing common strengths and challenges across a wide variety of classrooms, teachers, and children. Training raters to be able to perceive diverse classrooms in consistent, objective ways is a first step in pursuing that goal.

# References

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J. L., & Hsiung, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review, 22*(2), 207.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, D.C.: AERA.

Arnett, J. (1989). Caregivers in day-care centers: Does training matter. *Journal of Applied Developmental Psychology, 10*(4), 541-552.

Barnett, W. S., Epstein, D. J., Friedman, A. H., Boyd, J. S., & Hustedt, J. T. (2008). *The state of preschool 2008: State preschool yearbook*. New Brunswick, NJ: National Institute for Early Education Research, Rutgers University.

Bogard, K., & Takanishi, R. (2005). PK-3: An aligned and coordinated approach to education for children 3 to 8 years old. *Social Policy Report, 19*(3), 1-21.

Boyd, D., Goldhaber, D., Lankford, H., & Wyckoff, J. (2007). The Effect of Certification and Preparation on Teacher Quality. *Future of Children, 17*(1), 45-68.

Bradley, R. H., Caldwell, B. M., & Corwyn, R. F. (2003). The Child Care HOME Inventories: assessing the quality of family child care homes. *Early Childhood Research Quarterly, 18*(3), 294-309.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brooks-Gunn, J. (2003). Do you believe in magic? What we can expect from early childhood intervention programs. *Social Policy Report, 17*(1), 1-14.

Bryant, D. M. (2000). *Validating North Carolina's 5-star child care licensing system*. Chapel Hill, NC: Frank Porter Graham Child Development Center.

Bryant, D. M., Clifford, R. M., & Peisner, E. S. (1991). Best Practices for Beginners: Developmental Appropriateness in Kindergarten. *American Educational Research Journal, 28*(4), 783.

Burchinal, M., Howes, C., Pianta, R. C., Bryant, D., Early, D. M., Clifford, R. M., et al. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher–child interactions and instruction. *Applied Developmental Science, 12*(3), 140-153.

Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2009). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly*.

Cameron, C. E., Connor, C. M. D., Morrison, F. J., & Jewkes, A. M. (2008). Effects of classroom organization on letter–word reading in first grade. *Journal of School Psychology, 46*(2), 173-192.

Carrow-Woolfolk, E. (1995). *Oral and Written Language Scales (OWLS)*. Circle Pines, MN: American Guidance Service.

Clifford, R. M., Barbarin, O., Chang, F., Early, D., Bryant, D., Howes, C., et al. (2005). What is Pre-Kindergarten? Characteristics of Public Pre-Kindergarten Programs. *Applied Developmental Science, 9*(3), 126-143.

Clotfelter, C. T., Ladd, H. F., Vigdor, J., & Wheeler, J. (2006). High-poverty schools and the distribution of teachers and principals. *North Carolina Law Review, 85*, 1345.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources, 41*(4), 778.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review, 26*(6), 673-682.

Colvin, G., Flannery, K. B., Sugai, G., & Monegan, J. (2009). Using Observational Data to Provide Performance Feedback to Teachers: A High School Case Study. *Preventing School Failure, 53*(2), 95-104.

Connor, C. M. D., Morrison, F. J., Fishman, B. J., Ponitz, C. C., Glasney, S., Underwood, P. S., et al. (2009). The ISI Classroom Observation System: Examining the Literacy Instruction Provided to Individual Students. *Educational Researcher, 38*(2), 85.

Connor, C. M. D., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., et al. (2009). Individualizing student instruction precisely: Effects of child by instruction interactions on first graders’ literacy development. *Child Development, 80*(1), 77–100.

Connor, C. M. D., Son, S. H., Hindman, A. H., & Morrison, F. J. (2005). Teacher qualifications, classroom practices, family characteristics, and preschool experience: Complex effects on first graders' vocabulary and early reading outcomes. *Journal of School Psychology, 43*(4), 343-375.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Curby, T. W., Brock, L. L., & Hamre, B. K. (2009). The role of consistency in preschool teacher-child interactions. *Manuscript under review*.

Curby, T. W., LoCasale-Crouch, J., Konold, T. R., Pianta, R. C., Howes, C., Burchinal, M., et al. (2009). The Relations of Observed Pre-K Classroom Quality Profiles to Children's Achievement and Social Competence. *Early Education & Development, 20*(2), 346-372.

Currie, J., & Thomas, D. (2000). School quality and the longer-term effects of Head Start. *The Journal of Human Resources, 35*(4), 755-774.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Davidson, M. R., Fields, M. K., & Yang, J. (2009). A randomized trial study of a preschool literacy curriculum: The importance of implementation. *Journal of Research on Educational Effectiveness, 2*(3), 177-208.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association, 94*(448), 1053-1062.

Dickinson, D. K., & Caswell, L. (2007). Building support for language and early literacy in preschool classrooms through in-service professional development: Effects of the Literacy Environment Enrichment Program (LEEP). *Early Childhood Research Quarterly, 22*(2), 243-260.

Domitrovich, C. E., Greenberg, m. T., Kusche, C., & Cortes, R. (2004). *The preschool PATHS curriculum*. State College, PA: Pennsylvania State University.

Downer, J. T., & Hamre, B. K. (2010). *Beliefs about intentional instruction*: Unpublished measure.

Downey, C. J., English, F. W., Steffy, B. E., Poston Jr, W. K., & Frase, L. E. (2004). *The three-minute classroom walk-through: Changing school supervisory practice one teacher at a time*. Thousand Oaks, CA: Corwin Press.

Duncan, A. (2009). *The early learning challenge: Raising the bar*. Retrieved from http://www.ed.gov/news/speeches/2009/11/11182009.html

Dunn, L. M., & Dunn, L. M. (1997). Peabody picture vocabulary test-revised. Circle Pines MN. *American Guidance Service.*

Early, D. M., Maxwell, K. L., Burchinal, M., Alva, S., Bender, R. H., Bryant, D., et al. (2007). Teachers' education, classroom quality, and young children's academic skills: Results from seven studies of preschool programs. *Child Development, 78*(2), 558-580.

Emmer, E. T., & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist, 36*(2), 103-112.

Entwisle, D. R., & Alexander, K. L. (1993). Entry into school: The beginning school transition and educational stratification in the United States. *Annual Review of Sociology, 19*, 401-423.

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *Journal of Educational Research, 94*(5), 275-282.

Feng, W. W., Jun, Y., & Xu, R. (2006). *A method/macro based on propensity score and Mahalanobis distance to reduce bias in treatment comparison in observational study*. Paper presented at the SAS Conference: PharmaSUG 2006. from http://www.lexjansen.com/pharmasug/2006/publichealthresearch/pr05.pdf

Frank Porter Graham Child Development Institute. (2003). Roadmaps to Quality. *Early Developments, 7*(2), 18-19.

Frank Porter Graham Child Development Institute. (2009). *Levels of training on the environment rating scales*. Retrieved from http://www.fpg.unc.edu/~ecers/training_levels.htm

Gates, B. (2009). *2009 annual letter*: Bill and Melinda Gates Foundation. http://www.gatesfoundation.org/annual-letter/Pages/2009-bill-gates-annual-letter.aspx

Goodson, B. D., Layzer, J. I., & Layzer, C. J. (2005). *Quality of early childhood care settings: Caregiver rating scale (QUEST)*. Cambridge, MA: Abt Associates Inc.

Halle, T., & Vick, J. (2007). *Quality in early childhood care and education settings: A compendium of measures*. Washington, DC: Child Trends for the Office of Planning, Research and Evaluation, Administration for Children and Families, US Department of Health and Human Services.

Hamre, B. K., LoCasale-Crouch, J., & Pianta, R. C. (2007). Formative assessment of classrooms: Using classroom observations to improve implementation quality. In L. M. Justice, C. Vukelich & W. H. Teale (Eds.), *Achieving Excellence in Preschool Literacy Instruction*: The Guilford Press.

Hamre, B. K., & Pianta, R. C. (2001). Early teacher-child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development, 72*(2), 625-638.

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*(5), 949-967.

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2008). Building a science of classrooms: Application of the CLASS framework in over 4,000 US early childhood and elementary classrooms. Retrieved December 1, 2008, from http://www.fcd-us.org/resources/resources_show.htm?doc_id=507559

Harms, T., & Clifford, R. (1989). *The Family Day Care Rating Scale*. New York: Teachers College Press.

Harms, T., & Clifford, R. M. (1980). *The Early Childhood Environment Rating Scale*. New York: Teachers College Press.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *The Early Childhood Environment Rating Scale: Revised edition*. New York: Teachers College Press.

Harms, T., Cryer, D., & Clifford, R. M. (1990). *Infant/Toddler Environment Rating Scale*. New York: Teachers College Press.

Harms, T., Jacobs, E., & White, D. (1996). *School Age Care Environment Rating Scale*. New York: Teachers College Press.

Harris, D. N., & Sass, T. R. (2006). Value-Added Models and the Measurement of Teacher Quality. *Preliminary Draft, Unpublished manuscript, Florida State University, April*.

Hightower, A. D., Work, W. C., Cowen, E. L., Lotyczewski, B. S., Spinell, A. P., Guare, J. C., et al. (1986). The Teacher-Child Rating Scale: A brief objective measure of elementary children's school problem behaviors and competencies. *School Psychology Review, 15,* 393-409.

Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review, 34*(4), 507.

Hintze, J. M., & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review, 33*(2), 258-271.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., et al. (2008). Ready to learn? Children's pre-academic achievement in pre-Kindergarten programs. *Early Childhood Research Quarterly, 23*(1), 27-50.

Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*(4), 403-424.

Hsieh, W. Y., Hemmeter, M. L., McCollum, J. A., & Ostrosky, M. M. (2009). Using coaching to increase preschool teachers' use of emergent literacy teaching strategies. *Early Childhood Research Quarterly*.

Jepsen, C. (2005). Teacher characteristics and student achievement: evidence from teacher surveys. *Journal of Urban Economics, 57*(2), 302-319.

Johnson, R. L., Penny, J. A., & Gordon, B. (2008). *Assessing performance: Designing, scoring, and validating performance tasks*. New York: The Guilford Press.

Justice, L. M., Pullen, P. C., Hall, A., & Pianta, R. C. (2003). *MyTeachingPartner language and literacy curriculum*. Charlottesville, VA: University of Virginia Center for Advanced Study of Teaching and Learning.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2007). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*.

La Paro, K. M., Hamre, B. K., LoCasale-Crouch, J., Pianta, R. C., Bryant, D. M., Early, D. M., et al. (in press). Quality in kindergarten classrooms: Observational evidence for the need to increase children's learning opportunities in early education classrooms. *Early Education & Development*.

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The Classroom Assessment Scoring System: Findings from the prekindergarten year. *Elementary School Journal, 104*(5), 409-426.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis, 24*(1), 37.

Leff, S. S., & Lakin, R. (2005). Playground-based observational systems: A review and implications for practitioners and researchers. *School Psychology Review, 34*(4), 474.

LoCasale-Crouch, J., Downer, J. T., & Hamre, B. K. (2010). *Assessing teacher beliefs about intentional instruction*: Manuscript in preparation.

LoCasale-Crouch, J., Konold, T., Pianta, R., Howes, C., Burchinal, M., Bryant, D., et al. (2007). Observed classroom quality profiles in state-funded pre-kindergarten programs and associations with teacher, program, and classroom characteristics. *Early Childhood Research Quarterly, 22*(1), 3-17.

Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2007). Does prekindergarten improve school preparation and performance? *Economics of Education Review, 26*(1), 33.

Magnuson, K. A., & Waldfogel, J. (2005). Early childhood care and education: Effects on ethnic and racial gaps in school readiness. *The Future of Children, 15*(1), 169-196.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O., Bryant, D., et al. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development, 79*(3), 732-749.

McCaffrey, D. F., Sass, T. R., & Lockwood, J. R. (2008). The intertemporal stability of teacher effect estimates. *Preliminary Draft, Unpublished manuscript, June*.

Medina, J. (2009, September 1). A 2-year study to learn what makes teachers good. *The New York Times*. Retrieved from http://cityroom.blogs.nytimes.com/2009/09/01/a-2-year-study-to-learn-what-makes-teachers-good/

Merrell, K. W. (1999). *Behavioral, social, and emotional assessment of children and adolescents*. Mahwah, N.J.: Lawrence Erlbaum.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics, 13*(2), 151-161.

Meyer, J. P., Henry, A. E., & Mashburn, A. J. (2009). *The influence of occasion on the reliability of classroom observations: An application of multivariate generalizability theory*. Paper presented at the Northeastern Educational Research Association, Rocky Mount, CT.

Michalopoulos, C., Bloom, H. S., & Hill, C. J. (2004). Can Propensity-Score Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs? *Review of Economics and Statistics, 86*(1), 156-179.

Montgomery, D. (2002). *Helping teachers develop through classroom observation*. London: David Fulton Publishers Ltd.

Moon, T. R., & Hughes, K. R. (2005). Training and scoring issues involved in large-scale writing assessments. *Educational Measurement: Issues and Practice, 21*(2), 15-19.

National Association for the Education of Young Children. (2009). Developmentally appropriate practice in early childhood programs serving children from birth through age 8.   Retrieved February 24, 2009, from http://www.naeyc.org/about/positions.asp

National Association of Child Care Resource and Referral Agencies. (2009). *Comparison of Quality Rating and Improvement Systems (QRIS) with Department of Defense standards for quality* (No. 724-0714). Retrieved from http://www.naccrra.org/publications/naccrra-publications/comparison-qris-dod-quality-standards-2009.php

National Center for Education Statistics. (1994). *School and Staffing Survey 1993-1994; Principal's Survey*. Washington, DC: U.S. Department of Education.

National Child Care Information and Technical Assistance Center. (2007). *Child Care Bulletin Issue 32*. Fairfax, VA: Child Care Bureau, Administration for Children and Families, U.S. Department of Health and Human Services.

National Child Care Information and Technical Assistance Center. (2009). *Child care and development fund report of state and territory plans FY 2008-2009*. Retrieved from http://nccic.acf.hhs.gov/pubs/stateplan2008-09/index.html

National Council on Teacher Quality. (2007). State teacher policy yearbook: Progress on teacher quality.   Retrieved January 1, 2008, from http://www.nctq.org/stpy/

Neuman, S. B., & Cunningham, L. (2009). The impact of professional development and coaching on early language and literacy instructional practices. *American Educational Research Journal, 46*(2), 532.

Neuman, S. B., Koh, S., & Dwyer, J. (2008). CHELLO: The Child/Home Environmental Language and Literacy Observation. *Early Childhood Research Quarterly, 23*(2), 159-172.

NICHD Early Child Care Research Network. (2002). The relation of global first-grade classroom environment to structural classroom features and teacher and student behaviors. *Elementary School Journal, 102*(5), 367-387.

NICHD Early Child Care Research Network. (2005). A day in third grade: A large-scale study of classroom quality and teacher and student behavior. *Elementary School Journal, 105*(3).

NICHD Early Child Care Research Network, & Duncan, G. J. (2003). Modeling the impacts of child care quality on children's preschool cognitive development. *Child Development, 74*(5), 1454-1475.

NICHD ECCRN. (2000). The relation of child care to cognitive and language development. *Child Development, 71*(4), 960-980.

NICHD ECCRN. (2002a). Child-Care Structure Process Outcome: Direct and Indirect Effects of Child-Care Quality on Young Children's Development. *Psychological Science, 13*(3), 199-206.

NICHD ECCRN. (2002b). The relation of global first-grade classroom environment to structural classroom features and teacher and student behaviors. *Elementary School Journal, 102*(5), 367-387.

NICHD ECCRN. (2005). A day in third grade: A large-scale study of classroom quality and teacher and student behavior. *Elementary School Journal, 105*(3).

NICHD ECCRN. (2006). Child-Care Effect Sizes for the NICHD Study of Early Child Care and Youth Development. *American Psychologist, 61*(2), 99-116.

NICHD ECCRN, & Duncan, G. J. (2003). Modeling the impacts of child care quality on children's preschool cognitive development. *Child Development, 74*(5), 1454-1475.

Norris, D. J., Dunn, L., & Eckert, L. (2003). *Reaching for the Stars: Center validation study final report*: Early Childhood Collaborative of Oklahoma.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large Are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78*(1), 33.

Peisner-Feinberg, E. S., Burchinal, M. R., Clifford, R. M., Culkin, M. L., Howes, C., Kagan, S. L., et al. (2001). The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child Development, 72*(5), 1534-1553.

Penny, J. A. (2003). Reading high stakes writing samples: My life as a reader. *Assessing Writing, 8*(3), 192-215.

Pianta, R. C. (2001). Student Teacher Relationship Scale. Lutz, FL: Psychological Assessment Resources, Inc.

Pianta, R. C., & Allen, J. P. (2008). Building capacity for positive youth development in secondary school classrooms: Changing teachers' interactions with students. In *Toward Positive Youth Development: Transforming Schools and Community Programs*.

Pianta, R. C., Belsky, J., Houts, R., Morrison, F., & NICHD ECCRN. (2007). Observed classroom experiences in elementary school: A day in fifth grade and stability from grades 1 to 3. Manuscript submitted for publication.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109.

Pianta, R. C., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., et al. (2005). Features of pre-kindergarten programs, classrooms, and teachers: Do they predict observed classroom quality and child-teacher interactions? *Applied Developmental Science, 9*(3), 144-159.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008a). *Classroom Assessment Scoring System*. Baltimore, MD: Brookes Publishing.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008b). Technical appendix. In *Classroom Assessment Scoring System* (pp. 87-106). Baltimore, MD: Brookes Publishing.

Pianta, R. C., La Paro, K. M., Payne, C., Cox, M. J., & Bradley, R. (2002). The relation of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes. *The Elementary School Journal, 102*(3), 225(215).

Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher–child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly*.

Preschool Curriculum Evaluation Research Consortium. (2008). *Effects of preschool curriculum programs on school readiness* (NCER No. 2008-2009). Washington, DC: Institute of Education Sciences, National Center for Education Research. Retrieved from http://ncer.ed.gov

Pressley, M., Wharton-McDonald, R., Allington, R., Collins Block, C., Morrow, L., Tracey, D., et al. (2001). A study of effective first-grade literacy instruction. *Scientific Studies of Reading, 5*(1), 35-58.

Ramey, C. T., Campbell, F. A., Burchinal, M., Skinner, M. L., Gardner, D. M., & Ramey, S. L. (2000). Persistent effects of early childhood education on high-risk children and their mothers. *Applied Developmental Science, 4*(1), 2-14.

Raudenbush, S. W. (2005). Learning from Attempts to Improve Schooling: The Contribution of Methodological Diversity. *Educational Researcher, 34*(5), 25.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. London: Sage Publications.

Raudenbush, S. W., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness, 1*(2), 138-154.

Raudenbush, S. W., & Sampson, R. J. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology, 29*, 1-41.

Reynolds, A. J., Temple, J. A., Robertson, D. L., & Mann, E. A. (2001). Long-term effects of an early childhood intervention on educational achievement and juvenile arrest: A 15-year follow-up of low-income children in public schools. *JAMA: Journal of the American Medical Association, 285*(18), 2339-2346.

Rimm-Kaufman, S. E., Curby, T. W., Grimm, K. J., Nathanson, L., & Brock, L. L. (2009). The contribution of children’s self-regulation and classroom quality to children’s adaptive behaviors in the kindergarten classroom. *Developmental Psychology, 45*(4), 958-972.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417-458.

Robelen, E. W. (2008, November 12). Gates' new approach gets good reviews. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2008/11/12/13gatesreact.h28.html

Robelen, E. W. (2009, January 22). Gates gives $22 million in grants. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2009/01/22/19gates.h28.html

Rock, D. A., & Stenner, A. J. (2005). Assessment issues in the testing of children at school entry. *The Future of Children, 15*(1), 15(20).

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247-252.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33-38.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*(3-4), 169-188.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95*(450), 573-585.

Schaefer, E. S., & Edgerton, M. (1985). Parent and child correlates of parental modernity. *Parental belief systems: The psychological consequences for children,* 287–318.

Schafer, W. D., Gagné, P., & Lissitz, R. W. (2005). Resistance to Confounding Style and Content in Scoring Constructed-Response Items. *Educational Measurement: Issues and Practice, 24*(2), 22-28.

Schweinhart, L. J., & Weikart, D. P. (1997). The High/Scope Preschool Curriculum Comparison Study through age 23. *Early Childhood Research Quarterly, 12*(2), 117.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can randomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association, 103*(484), 1334-1356.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*: Houghton Mifflin.

Simon, A., & Boyer, E. G. (1969). *Mirrors for Behavior, An Anthology of Classroom Observation Instruments*. Philadelphia, PA: Research for Better Schools, Inc.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23*(4), 323.

Smith, J. (2000). *A critical survey of empirical methods for evaluating active labor market policies*: Department of Economics, University of Western Ontario.

Smith, M., Dickinson, D. K., Sangeorge, A., & Anastasopoulos, L. (2002). Early language and literacy classroom observation (ELLCO) toolkit: Baltimore: Brookes.

Stipek, D., & Byler, P. (2004). The early childhood classroom observation measure. *Early Childhood Research Quarterly, 19*(3), 375-397.

Stoney, L. (2004). *Financing Quality Rating Systems: Lessons learned*: Alliance for Early Childhood Finance for United Way of America Success by 6.

Stuhlman, M., Curby, T. W., Grimm, K. J., Mashburn, A., Chomat-Mooney, L., Hamre, B. K., et al. (2009). Within-day variability in third and fifth grade in classroom interaction quality: Implications for children's experience and conducting classroom observation. *Manuscript in preparation*.

Stuhlman, M., & Pianta, R. C. (2009). Profiles of educational quality in first grade. *The Elementary School Journal, 109*(4), 323-342.

Teachstone. (2009). *CLASS regional training*. Retrieved from http://www.teachstone.org/regional_training.php

Tout, K., Starr, R., & Cleveland, J. (2008). *Evaluation of Parent Aware: Minnesota's Quality Rating System pilot*. Minneapolis, MN: Minnesota Early Learning Foundation Research Consortium.

Tout, K., Zaslow, M., Halle, T., & Forry, N. (2009). *Issues for the next decade of Quality Rating and Improvement Systems* (Issue Brief No. 3). Washington, DC: Prepared by Child Trends for the Office of Planning, Research, and Evaluation, Administration for Children and Families, US Department of Health and Human Services.

U.S. Department of Health and Human Services, Administration for Children and Families, & Office of Head Start. (2008). *Classroom Assessment Scoring System* (Information Memorandum No. ACF--IM-HS-08-11). Washington, DC: Brown, Patricia E.

Wachs, T. D., Gurkas, P., & Kontos, S. (2004). Predictors of preschool children's compliance behavior in early childhood classroom settings. *Journal of Applied Developmental Psychology, 25*(4), 439-457.

Wasik, B. A., Bond, M. A., & Hindman, A. (2006). The effects of a language and literacy intervention on Head Start children and teachers. *Journal of Educational Psychology, 98*(1), 63.

Wiley, C., Good, T., & McCaslin, M. (2008). Comprehensive school reform instructional practices throughout a school year: The role of subject matter, grade level, and time of year. *The Teachers College Record, 110*(11), 2361-2388.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III: Tests of Achievement*. Itasca, ILL: Riverside Publishing.

Zaslow, M., Tout, K., Halle, T., & Forry, N. (2009). *Multiple purposes for measuring quality in early childhood settings: Implications for collecting and communicating information on quality* (Issue Brief No. 2). Washington, DC: Prepared by Child Trends for the Office of Planning, Research, and Evaluation, Administration for Children and Families, US Department of Health and Human Services.

Zellman, G. L., & Perlman, M. (2008). *Child-care Quality Rating and Improvement Systems in five pioneer states: Implementation issues and lessons learned*. Arlington, VA: RAND Corporation.

Zellman, G. L., Perlman, M., Le, V.-N., & Setodji, C. M. (2008). *Assessing the validity of the Qualistar Early Learning Quality Rating and Improvement System as a tool for improving child-care quality*. Arlington, VA: RAND Corporation.

Zepeda, S. J. (2008). *The instructional leader's guide to informal classroom observations*. Larchmont, NY: Eye on Education.

Zill, N., Resnick, G., Kim, K., McKey, R. H., Clark, C., Pai-Samant, S., et al. (2001). *Head Start FACES: Longitudinal Findings on Program Performance. Third Progress Report.* http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED453969&site=ehost-live

Zill, N., Sorongon, A., Kim, K., Clark, C., & Woolverton, M. (2006). *FACES 2003 research brief*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families. Retrieved from www.acf.hhs.gov/programs/opre/hs/faces/index.html

Table 1

*Descriptive Statistics for Adjacent Calibration Scores, Percent Within-1 of Master-Code*

| Outcome | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|
| Full Sample | | | | | |
| Overall | 2093 | 82.94 | 10.55 | 20.00 | 100.00 |
| Positive Climate | 2093 | 88.73 | 18.96 | 0.00 | 100.00 |
| Negative Climate | 2093 | 93.31 | 14.25 | 0.00 | 100.00 |
| Teacher Sensitivity | 2093 | 85.59 | 20.43 | 0.00 | 100.00 |
| Regard for Student Perspectives | 2093 | 82.84 | 22.19 | 0.00 | 100.00 |
| Behavior Management | 2093 | 86.04 | 19.32 | 0.00 | 100.00 |
| Productivity | 2093 | 82.51 | 20.52 | 0.00 | 100.00 |
| Instructional Learning Formats | 2093 | 89.69 | 18.62 | 0.00 | 100.00 |
| Concept Development | 2093 | 66.54 | 25.55 | 0.00 | 100.00 |
| Quality of Feedback | 2093 | 80.38 | 26.42 | 0.00 | 100.00 |
| Language Modeling | 2093 | 72.48 | 30.77 | 0.00 | 100.00 |
| Survey Sample | | | | | |
| Overall | 704 | 84.50 | 9.91 | 43.00 | 100.00 |
| Positive Climate | 704 | 89.72 | 17.89 | 0.00 | 100.00 |
| Negative Climate | 704 | 94.52 | 13.14 | 33.33 | 100.00 |
| Teacher Sensitivity | 704 | 87.00 | 19.02 | 33.33 | 100.00 |
| Regard for Student Perspectives | 704 | 82.98 | 22.07 | 0.00 | 100.00 |
| Behavior Management | 704 | 87.81 | 18.56 | 0.00 | 100.00 |
| Productivity | 704 | 82.80 | 20.34 | 0.00 | 100.00 |
| Instructional Learning Formats | 704 | 90.97 | 18.01 | 0.00 | 100.00 |
| Concept Development | 704 | 69.46 | 24.66 | 0.00 | 100.00 |
| Quality of Feedback | 704 | 83.46 | 24.24 | 0.00 | 100.00 |
| Language Modeling | 704 | 75.87 | 29.87 | 0.00 | 100.00 |

Table 2

*Descriptive Statistics for Valence Distance Scores*

| Outcome | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|
| | | Full Sample | | | |
| Overall | 2093 | 0.31 | 0.39 | -1.00 | 1.93 |
| Positive Climate | 2093 | -0.07 | 0.61 | -2.67 | 2.00 |
| Negative Climate | 2093 | 0.31 | 0.45 | -0.33 | 3.33 |
| Teacher Sensitivity | 2093 | -0.07 | 0.59 | -2.33 | 2.33 |
| Regard for Student Perspectives | 2093 | 0.45 | 0.63 | -1.67 | 2.48 |
| Behavior Management | 2093 | -0.25 | 0.57 | -3.33 | 1.67 |
| Productivity | 2093 | -0.11 | 0.55 | -2.33 | 1.67 |
| Instructional Learning Formats | 2093 | 0.08 | 0.61 | -2.33 | 2.33 |
| Concept Development | 2093 | 1.05 | 0.69 | -1.00 | 3.67 |
| Quality of Feedback | 2093 | 0.71 | 0.72 | -1.33 | 3.67 |
| Language Modeling | 2093 | 0.96 | 0.75 | -1.33 | 3.67 |
| | | Survey Sample | | | |
| Overall | 704 | 0.27 | 0.37 | -0.90 | 1.73 |
| Positive Climate | 704 | -0.07 | 0.61 | -2.00 | 2.00 |
| Negative Climate | 704 | 0.29 | 0.42 | -0.33 | 2.00 |
| Teacher Sensitivity | 704 | -0.09 | 0.57 | -2.00 | 1.67 |
| Regard for Student Perspectives | 704 | 0.43 | 0.60 | -1.33 | 2.00 |
| Behavior Management | 704 | -0.26 | 0.56 | -3.33 | 1.67 |
| Productivity | 704 | -0.14 | 0.54 | -1.67 | 1.67 |
| Instructional Learning Formats | 704 | 0.08 | 0.60 | -2.00 | 2.00 |
| Concept Development | 704 | 0.96 | 0.68 | -1.00 | 3.33 |
| Quality of Feedback | 704 | 0.63 | 0.67 | -1.00 | 3.00 |
| Language Modeling | 704 | 0.88 | 0.73 | -1.00 | 3.33 |

Table 3

*Descriptive Statistics for Absolute Distance Scores*

| Outcome | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|
| | | Full Sample | | | |
| Overall | 2093 | 0.39 | 0.31 | 0.00 | 1.93 |
| Positive Climate | 2093 | 0.47 | 0.39 | 0.00 | 2.67 |
| Negative Climate | 2093 | 0.40 | 0.37 | 0.00 | 3.33 |
| Teacher Sensitivity | 2093 | 0.46 | 0.38 | 0.00 | 2.33 |
| Regard for Student | | | | | |
| Perspectives | 2093 | 0.61 | 0.47 | 0.00 | 2.48 |
| Behavior Management | 2093 | 0.49 | 0.39 | 0.00 | 3.33 |
| Productivity | 2093 | 0.43 | 0.36 | 0.00 | 2.33 |
| Instructional Learning Formats | 2093 | 0.48 | 0.39 | 0.00 | 2.33 |
| Concept Development | 2093 | 1.07 | 0.66 | 0.00 | 3.67 |
| Quality of Feedback | 2093 | 0.79 | 0.63 | 0.00 | 3.67 |
| Language Modeling | 2093 | 1.00 | 0.69 | 0.00 | 3.67 |
| | | Survey Sample | | | |
| Overall | 704 | 0.36 | 0.29 | 0.00 | 1.73 |
| Positive Climate | 704 | 0.47 | 0.39 | 0.00 | 2.00 |
| Negative Climate | 704 | 0.37 | 0.35 | 0.00 | 2.00 |
| Teacher Sensitivity | 704 | 0.44 | 0.37 | 0.00 | 2.00 |
| Regard for Student | | | | | |
| Perspectives | 704 | 0.59 | 0.44 | 0.00 | 2.00 |
| Behavior Management | 704 | 0.49 | 0.37 | 0.00 | 3.33 |
| Productivity | 704 | 0.43 | 0.36 | 0.00 | 1.67 |
| Instructional Learning Formats | 704 | 0.46 | 0.39 | 0.00 | 2.00 |
| Concept Development | 704 | 0.98 | 0.65 | 0.00 | 3.33 |
| Quality of Feedback | 704 | 0.72 | 0.57 | 0.00 | 3.00 |
| Language Modeling | 704 | 0.94 | 0.66 | 0.00 | 3.33 |

Table 4a

*Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Predictors of Absolute Distance, Full Sample*

| | Overall | | | | Positive Climate | | | |
|---|---|---|---|---|---|---|---|---|
| Parameter | Model 1 | | Model 2 | | Model 1 | | Model 2 | |
| | Fixed effects | | | | | | | |
| Intercept | 0.394 *** | (0.011) | 0.400 *** | (0.083) | 0.476 *** | (0.010) | 0.686 *** | (0.082) |
| Level-2 (Session) | | | | | | | | |
| Trainer 2 | | | -0.030 *** | (0.134) | | | 0.011 | (0.130) |
| Trainer 3 | | | 0.049 | (0.111) | | | -0.248 * | (0.106) |
| Trainer 4 | | | 0.140 | (0.134) | | | -0.083 | (0.127) |
| Trainer 5 | | | 0.017 | (0.079) | | | -0.227 ** | (0.078) |
| Trainer 6 | | | -0.021 | (0.091) | | | -0.152 † | (0.089) |
| Trainer 7 | | | 0.017 | (0.078) | | | -0.105 | (0.078) |
| Trainer 8 | | | -0.031 | (0.092) | | | -0.105 | (0.090) |
| Trainer 9 | | | -0.165 | (0.113) | | | -0.133 | (0.109) |
| Trainer 10 | | | 0.215 † | (0.119) | | | -0.172 | (0.122) |
| Trainer 11 | | | -0.013 | (0.091) | | | -0.221 | (0.089) |
| Trainer 12 | | | 0.064 | (0.136) | | | -0.141 | (0.131) |
| Trainer 13 | | | -0.059 | (0.089) | | | -0.156 † | (0.088) |
| Trainer 14 | | | 0.090 | (0.103) | | | -0.203 * | (0.100) |
| Trainer 15 | | | 0.080 | (0.083) | | | -0.202 * | (0.082) |
| Trainer 16 | | | 0.059 | (0.101) | | | -0.187 † | (0.099) |
| Trainer 17 | | | 0.018 | (0.135) | | | -0.058 | (0.130) |
| Trainer 18 | | | 0.105 | (0.109) | | | -0.095 | (0.105) |
| Trainer 19 | | | -0.064 | (0.081) | | | -0.197 * | (0.080) |
| Trainer 20 | | | 0.117 | (0.093) | | | -0.185 * | (0.090) |
| Trainer 21 | | | -0.037 | (0.079) | | | -0.142 † | (0.078) |
| Trainer 22 | | | 0.122 | (0.086) | | | -0.200 * | (0.084) |
| Trainer 23 | | | -0.051 | (0.104) | | | -0.224 * | (0.100) |
| Trainer 24 | | | -0.067 | (0.097) | | | -0.234 * | (0.095) |
| Trainer 25 | | | -0.011 | (0.092) | | | -0.167 † | (0.090) |
| Days since 11/1/08 | | | -0.0001 | (0.0002) | | | -0.0003 † | (0.0002) |
| Number of raters | | | -0.002 | (0.003) | | | -0.004 | (0.003) |
| | Random effects | | | | | | | |
| Level-2 Intercept (uj) | 0.010 *** | (0.002) | 0.008 *** | (0.002) | 0.004 * | (0.002) | 0.003 † | (0.002) |
| Level-1 Residual (rij) | 0.086 *** | (0.003) | 0.086 *** | (0.003) | 0.149 *** | (0.005) | 0.148 *** | (0.005) |

*Note*. Standard errors are in parentheses.

†*p* <.10. **p* <.05. ***p* <.01. ****p* <.001.

Table 4b

*Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Predictors of Absolute Distance, Full Sample*

| | Negative Climate | | | | Teacher Sensitivity | | | | Regard for Student Perspectives | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Parameter | Model 1 | | Model 2 | | Model 1 | | Model 2 | | Model 1 | | Model 2 | |
| Fixed effects | | | | | | | | | | | | |
| Intercept | 0.403 *** | (0.012) | 0.569 *** | (0.090) | 0.461 *** | (0.010) | 0.540 *** | (0.075) | 0.614 *** | (0.016) | 0.594 *** | (0.122) |
| Level-2 (Session) | | | | | | | | | | | | |
| Trainer 2 | | | 0.185 *** | (0.145) | | | 0.298 * | (0.117) | | | -0.216 *** | (0.196) |
| Trainer 3 | | | -0.255 * | (0.119) | | | -0.103 | (0.095) | | | 0.175 | (0.162) |
| Trainer 4 | | | -0.040 | (0.143) | | | 0.174 | (0.114) | | | 0.011 | (0.195) |
| Trainer 5 | | | -0.086 | (0.086) | | | -0.104 | (0.071) | | | -0.050 | (0.116) |
| Trainer 6 | | | -0.126 | (0.098) | | | -0.064 | (0.081) | | | -0.107 | (0.133) |
| Trainer 7 | | | -0.125 | (0.085) | | | -0.031 | (0.071) | | | 0.126 | (0.115) |
| Trainer 8 | | | -0.018 | (0.099) | | | 0.012 | (0.082) | | | 0.001 | (0.134) |
| Trainer 9 | | | -0.005 | (0.121) | | | -0.003 | (0.098) | | | -0.182 | (0.165) |
| Trainer 10 | | | -0.199 | (0.130) | | | -0.112 | (0.113) | | | 0.318 † | (0.175) |
| Trainer 11 | | | -0.170 † | (0.098) | | | -0.112 | (0.082) | | | -0.106 | (0.133) |
| Trainer 12 | | | -0.217 | (0.146) | | | -0.082 | (0.118) | | | 0.210 | (0.198) |
| Trainer 13 | | | -0.125 | (0.097) | | | -0.018 | (0.080) | | | -0.082 | (0.131) |
| Trainer 14 | | | -0.069 | (0.111) | | | -0.045 | (0.093) | | | 0.023 | (0.151) |
| Trainer 15 | | | -0.108 | (0.090) | | | -0.062 | (0.075) | | | 0.039 | (0.122) |
| Trainer 16 | | | -0.177 | (0.109) | | | 0.015 | (0.089) | | | 0.120 | (0.148) |
| Trainer 17 | | | -0.258 † | (0.145) | | | -0.104 | (0.117) | | | -0.090 | (0.198) |
| Trainer 18 | | | -0.152 | (0.117) | | | 0.009 | (0.095) | | | 0.087 | (0.159) |
| Trainer 19 | | | -0.138 | (0.088) | | | -0.066 | (0.073) | | | -0.105 | (0.119) |
| Trainer 20 | | | -0.072 | (0.100) | | | -0.021 | (0.082) | | | 0.143 | (0.135) |
| Trainer 21 | | | -0.009 | (0.086) | | | -0.006 | (0.071) | | | -0.074 | (0.116) |
| Trainer 22 | | | -0.149 | (0.093) | | | -0.045 | (0.076) | | | 0.105 | (0.125) |
| Trainer 23 | | | 0.017 | (0.112) | | | -0.094 | (0.090) | | | -0.074 | (0.152) |
| Trainer 24 | | | -0.090 *** | (0.105) | | | -0.149 † | (0.086) | | | -0.052 | (0.142) |
| Trainer 25 | | | -0.222 * | (0.099) | | | -0.041 | (0.082) | | | -0.036 | (0.135) |
| Days since 11/1/08 | | | -0.0005 * | (0.0002) | | | -0.0003 | (0.0002) | | | 0.0001 | (0.0003) |
| Number of raters | | | 0.001 | (0.004) | | | -0.006 † | (0.003) | | | -0.003 | (0.005) |
| Random effects | | | | | | | | | | | | |
| Level-2 Intercept (uj) | 0.010 *** | (0.002) | 0.008 *** | (0.002) | 0.003 * | (0.002) | 0.001 | (0.001) | 0.020 *** | (0.004) | 0.016 *** | (0.004) |
| Level-1 Residual (rij) | 0.124 *** | (0.004) | 0.124 *** | (0.004) | 0.143 *** | (0.005) | 0.143 *** | (0.005) | 0.203 *** | (0.007) | 0.203 *** | (0.007) |

*Note*. Standard errors are in parentheses.

†*p* <.10. **p* <.05. ***p* <.01. ****p* <.001.

Table 4c

*Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Predictors of Absolute Distance, Full Sample*

| | Behavior Management | | | | Productivity | | | | Instructional Learning Formats | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Parameter | Model 1 | | Model 2 | | Model 1 | | Model 2 | | Model 1 | | Model 2 | |
| | | | | | Fixed effects | | | | | | | |
| Intercept | 0.492 *** | (0.011) | 0.510 *** | (0.087) | 0.433 *** | (0.010) | 0.416 *** | (0.085) | 0.478 *** | (0.011) | 0.606 *** | (0.087) |
| Level-2 (Session) | | | | | | | | | | | | |
| Trainer 2 | | | 0.194 *** | (0.137) | | | 0.001 *** | (0.135) | | | -0.070 *** | (0.137) |
| Trainer 3 | | | -0.002 | (0.112) | | | -0.094 | (0.111) | | | -0.050 | (0.112) |
| Trainer 4 | | | 0.080 | (0.135) | | | 0.070 | (0.134) | | | -0.134 | (0.135) |
| Trainer 5 | | | 0.024 | (0.082) | | | -0.023 | (0.081) | | | -0.170 * | (0.082) |
| Trainer 6 | | | -0.045 | (0.094) | | | -0.051 | (0.092) | | | -0.088 | (0.094) |
| Trainer 7 | | | 0.121 | (0.082) | | | 0.057 | (0.080) | | | 0.011 | (0.082) |
| Trainer 8 | | | -0.062 | (0.095) | | | 0.002 | (0.093) | | | -0.089 | (0.095) |
| Trainer 9 | | | 0.156 | (0.115) | | | 0.057 | (0.113) | | | -0.098 | (0.115) |
| Trainer 10 | | | 0.000 | (0.127) | | | -0.051 | (0.123) | | | 0.033 | (0.128) |
| Trainer 11 | | | 0.034 | (0.094) | | | -0.079 | (0.092) | | | -0.142 | (0.094) |
| Trainer 12 | | | -0.010 | (0.139) | | | -0.070 | (0.136) | | | -0.159 | (0.139) |
| Trainer 13 | | | 0.076 | (0.093) | | | -0.005 | (0.091) | | | -0.105 | (0.093) |
| Trainer 14 | | | 0.014 | (0.106) | | | 0.031 | (0.104) | | | -0.011 | (0.107) |
| Trainer 15 | | | -0.063 | (0.086) | | | -0.055 | (0.084) | | | -0.104 | (0.086) |
| Trainer 16 | | | -0.024 | (0.104) | | | -0.005 | (0.102) | | | -0.089 | (0.104) |
| Trainer 17 | | | 0.076 | (0.138) | | | 0.187 | (0.136) | | | -0.149 | (0.138) |
| Trainer 18 | | | 0.082 | (0.112) | | | 0.005 | (0.110) | | | 0.031 | (0.112) |
| Trainer 19 | | | 0.034 | (0.084) | | | -0.017 | (0.082) | | | -0.079 | (0.084) |
| Trainer 20 | | | 0.066 | (0.095) | | | 0.003 | (0.094) | | | -0.111 | (0.095) |
| Trainer 21 | | | 0.106 | (0.083) | | | 0.057 | (0.081) | | | -0.081 | (0.083) |
| Trainer 22 | | | 0.007 | (0.089) | | | 0.021 | (0.087) | | | -0.103 | (0.089) |
| Trainer 23 | | | 0.127 | (0.106) | | | 0.086 | (0.104) | | | -0.212 * | (0.106) |
| Trainer 24 | | | 0.103 | (0.100) | | | 0.118 | (0.098) | | | -0.183 † | (0.100) |
| Trainer 25 | | | -0.028 | (0.095) | | | 0.031 | (0.093) | | | -0.099 | (0.095) |
| Days since 11/1/08 | | | -0.0005 * | (0.0002) | | | 0.0001 | (0.0002) | | | -0.0003 | (0.0002) |
| Number of raters | | | -0.006 † | (0.003) | | | 0.002 | (0.003) | | | 0.000 | (0.003) |
| | | | | | Random effects | | | | | | | |
| Level-2 Intercept (uj) | 0.007 *** | (0.002) | 0.005 * | (0.002) | 0.006 *** | (0.002) | 0.006 ** | (0.002) | 0.005 ** | (0.002) | 0.004 * | (0.002) |
| Level-1 Residual (rij) | 0.143 *** | (0.005) | 0.143 *** | (0.005) | 0.123 *** | (0.004) | 0.123 *** | (0.004) | 0.151 *** | (0.005) | 0.150 *** | (0.005) |

*Note*. Standard errors are in parentheses.

†*p* <.10. **p* <.05. ***p* <.01. ****p* <.001.

Table 4d

*Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Predictors of Absolute Distance, Full Sample*

| | Concept Development | | Quality of Feedback | | Language Modeling | |
|---|---|---|---|---|---|---|
| Parameter | Model 1 | Model 2 | Model 1 | Model 2 | Model 1 | Model 2 |
| | | | Fixed effects | | | |
| Intercept | 1.072 *** (0.027) | 1.122 *** (0.180) | 0.794 *** (0.025) | 0.892 *** (0.172) | 1.002 *** (0.027) | 1.017 *** (0.186) |
| Level-2 (Session) | | | | | | |
| Trainer 2 | | 0.445 *** (0.290) | | 0.225 *** (0.276) | | 0.334 *** (0.299) |
| Trainer 3 | | 0.072 (0.240) | | 0.015 (0.228) | | 0.034 (0.247) |
| Trainer 4 | | 0.150 (0.290) | | 0.283 (0.276) | | 0.478 (0.299) |
| Trainer 5 | | 0.138 (0.171) | | 0.107 (0.162) | | 0.166 (0.176) |
| Trainer 6 | | 0.004 (0.197) | | -0.186 (0.187) | | -0.045 (0.203) |
| Trainer 7 | | -0.128 (0.169) | | -0.206 (0.161) | | -0.155 (0.175) |
| Trainer 8 | | -0.148 (0.198) | | -0.235 (0.188) | | -0.158 (0.204) |
| Trainer 9 | | -0.406 † (0.243) | | -0.437 † (0.232) | | -0.293 (0.251) |
| Trainer 10 | | 0.532 * (0.256) | | 0.170 (0.243) | | 0.203 (0.264) |
| Trainer 11 | | 0.282 (0.196) | | 0.125 (0.187) | | 0.225 (0.203) |
| Trainer 12 | | 0.011 (0.293) | | -0.222 (0.279) | | -0.035 (0.302) |
| Trainer 13 | | -0.122 (0.194) | | -0.172 (0.185) | | -0.101 (0.200) |
| Trainer 14 | | 0.143 (0.223) | | 0.035 (0.212) | | 0.219 (0.230) |
| Trainer 15 | | 0.209 (0.179) | | 0.149 (0.170) | | 0.189 (0.185) |
| Trainer 16 | | 0.045 (0.219) | | -0.079 (0.208) | | 0.080 (0.226) |
| Trainer 17 | | 0.223 (0.293) | | 0.130 (0.279) | | 0.299 (0.302) |
| Trainer 18 | | -0.107 (0.236) | | -0.100 (0.225) | | -0.117 (0.243) |
| Trainer 19 | | -0.055 (0.175) | | -0.078 (0.167) | | -0.087 (0.181) |
| Trainer 20 | | 0.271 (0.200) | | 0.230 (0.190) | | 0.333 (0.206) |
| Trainer 21 | | 0.036 (0.171) | | -0.063 (0.163) | | -0.128 (0.177) |
| Trainer 22 | | 0.389 * (0.185) | | 0.301 † (0.176) | | 0.399 * (0.191) |
| Trainer 23 | | 0.190 (0.224) | | 0.011 (0.214) | | 0.156 (0.231) |
| Trainer 24 | | 0.023 *** (0.210) | | -0.048 *** (0.200) | | -0.101 (0.217) |
| Trainer 25 | | 0.143 (0.199) | | 0.038 (0.189) | | 0.152 (0.205) |
| Days since 11/1/08 | | -0.001 * (0.000) | | -0.001 † (0.000) | | -0.001 (0.000) |
| Number of raters | | -0.004 (0.007) | | -0.006 (0.007) | | -0.003 (0.008) |
| | | | Random effects | | | |
| Level-2 Intercept (uj) | 0.062 *** (0.011) | 0.038 *** (0.009) | 0.053 *** (0.010) | 0.035 *** (0.008) | 0.065 *** (0.012) | 0.040 *** (0.009) |
| Level-1 Residual (rij) | 0.382 *** (0.012) | 0.381 *** (0.012) | 0.342 *** (0.011) | 0.342 *** (0.011) | 0.415 *** (0.013) | 0.415 *** (0.013) |

*Note*. Standard errors are in parentheses.

†*p* <.10. **p* <.05. ***p* <.01. ****p* <.001.

Table 5a

*Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Predictors of Absolute Distance, Survey Sample*

| | Overall | | | |
|---|---|---|---|---|
| Parameter | Model 1 | | Model 2 | |
| | Fixed effects | | | |
| Intercept | 0.364 *** | (0.016) | 0.173 | (0.352) |
| Level-1 (Rater) | | | | |
| Black | | | 0.093 ** | (0.032) |
| Latino | | | 0.067 † | (0.041) |
| Other race/ethnicity | | | 0.034 | (0.036) |
| MA or higher level of education | | | -0.009 | (0.022) |
| 3+ Courses in early childhood development | | | -0.040 † | (0.024) |
| Num. years supervisory experience | | | -0.001 | (0.002) |
| 6+ hours per week observing | | | 0.011 | (0.021) |
| Used ECERS-R in past year | | | -0.014 | (0.022) |
| Ideas About Children | | | 0.022 | (0.024) |
| Intentional Teaching Beliefs | | | -0.044 ** | (0.017) |
| Level-2 (Session) | | | | |
| Trainer 3 | | | -0.091 | (0.138) |
| Trainer 4 | | | 0.061 | (0.152) |
| Trainer 5 | | | 0.021 | (0.102) |
| Trainer 6 | | | -0.048 | (0.125) |
| Trainer 7 | | | -0.008 | (0.107) |
| Trainer 8 | | | -0.107 | (0.117) |
| Trainer 9 | | | -0.240 | (0.148) |
| Trainer 10 | | | -0.170 | (0.150) |
| Trainer 11 | | | -0.086 | (0.116) |
| Trainer 12 | | | 0.005 | (0.137) |
| Trainer 13 | | | -0.020 | (0.114) |
| Trainer 14 | | | (0.229) | (0.143) |
| Trainer 15 | | | 0.056 | (0.109) |
| Trainer 16 | | | -0.028 | (0.129) |
| Trainer 17 | | | -0.055 | (0.142) |
| Trainer 18 | | | 0.052 | (0.128) |
| Trainer 19 | | | -0.191 † | (0.116) |
| Trainer 20 | | | -0.046 | (0.116) |
| Trainer 21 | | | -0.100 | (0.102) |
| Trainer 22 | | | -0.011 | (0.112) |
| Trainer 23 | | | -0.200 | (0.145) |
| Trainer 24 | | | -0.125 | (0.111) |
| Trainer 25 | | | -0.156 | (0.119) |
| Days since 11/1/08 | | | -0.001 | (0.000) |
| Number of raters | | | 0.000 | (0.004) |
| Average Ideas About Children | | | 0.158 * | (0.079) |
| Average Intentional Teaching Beliefs | | | 0.035 | (0.063) |
| | Random effects | | | |
| Level-2 Intercept (uj) | 0.011 *** | (0.003) | 0.005 | (0.003) |
| Level-1 Residual (rij) | 0.070 *** | (0.004) | 0.068 *** | (0.004) |

*Note*. Standard errors are in parentheses.

†*p* <.10. **p* <.05. ***p* <.01. ****p* <.001.

Table 5b

*Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Predictors of Absolute Distance, Survey Sample*

| | Positive Climate | | | | | |
|---|---|---|---|---|---|---|
| Parameter | Model 1 | | | Model 2 | | |
| | Fixed effects | | | | | |
| Intercept | 0.476 | *** | (0.016) | 0.669 | | (0.465) |
| Level-1 (Rater) | | | | | | |
| Black | | | | 0.046 | | (0.046) |
| Latino | | | | -0.135 | * | (0.058) |
| Other race/ethnicity | | | | 0.002 | | (0.052) |
| MA or higher level of education | | | | 0.031 | | (0.032) |
| 3+ Courses in early childhood development | | | | 0.017 | | (0.035) |
| Num. years supervisory experience | | | | 0.002 | | (0.002) |
| 6+ hours per week observing | | | | -0.049 | | (0.031) |
| Used ECERS-R in past year | | | | -0.012 | | (0.032) |
| Ideas About Children | | | | 0.009 | | (0.035) |
| Intentional Teaching Beliefs | | | | 0.013 | | (0.025) |
| Level-2 (Session) | | | | | | |
| Trainer 3 | | | | -0.226 | | (0.172) |
| Trainer 4 | | | | 0.060 | | (0.196) |
| Trainer 5 | | | | -0.072 | | (0.133) |
| Trainer 6 | | | | -0.053 | | (0.165) |
| Trainer 7 | | | | 0.003 | | (0.140) |
| Trainer 8 | | | | 0.128 | | (0.152) |
| Trainer 9 | | | | -0.176 | | (0.188) |
| Trainer 10 | | | | 0.024 | | (0.195) |
| Trainer 11 | | | | 0.103 | | (0.151) |
| Trainer 12 | | | | -0.050 | | (0.172) |
| Trainer 13 | | | | -0.058 | | (0.147) |
| Trainer 14 | | | | 0.147 | | (0.180) |
| Trainer 15 | | | | -0.090 | | (0.143) |
| Trainer 16 | | | | 0.039 | | (0.165) |
| Trainer 17 | | | | 0.129 | | (0.178) |
| Trainer 18 | | | | 0.072 | | (0.165) |
| Trainer 19 | | | | 0.003 | | (0.149) |
| Trainer 20 | | | | -0.014 | | (0.150) |
| Trainer 21 | | | | -0.001 | | (0.133) |
| Trainer 22 | | | | -0.037 | | (0.144) |
| Trainer 23 | | | | -0.103 | | (0.182) |
| Trainer 24 | | | | -0.078 | | (0.143) |
| Trainer 25 | | | | 0.049 | | (0.154) |
| Days since 11/1/08 | | | | 0.000 | | (0.001) |
| Number of raters | | | | -0.001 | | (0.006) |
| Average Ideas About Children | | | | 0.014 | | (0.105) |
| Average Intentional Teaching Beliefs | | | | -0.082 | | (0.084) |
| | Random effects | | | | | |
| Level-2 Intercept (uj) | 0.003 | | (0.003) | 0.004 | | (0.005) |
| Level-1 Residual (rij) | 0.146 | *** | (0.008) | 0.145 | *** | (0.008) |

*Note*. Standard errors are in parentheses.
†*p* <.10. **p* <.05. ***p* <.01. ****p* <.001.

Table 5c

*Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Predictors of Absolute Distance, Survey Sample*

| | Negative Climate | | | |
|---|---|---|---|---|
| Parameter | Model 1 | | Model 2 | |
| | Fixed effects | | | |
| Intercept | 0.375 *** | (0.017) | 0.343 | (0.496) |
| Level-1 (Rater) | | | | |
| Black | | | 0.035 | (0.041) |
| Latino | | | 0.049 | (0.052) |
| Other race/ethnicity | | | -0.027 | (0.046) |
| MA or higher level of education | | | 0.031 | (0.028) |
| 3+ Courses in early childhood development | | | 0.058 † | (0.030) |
| Num. years supervisory experience | | | 0.003 † | (0.002) |
| 6+ hours per week observing | | | 0.031 | (0.027) |
| Used ECERS-R in past year | | | -0.059 * | (0.028) |
| Ideas About Children | | | 0.007 | (0.030) |
| Intentional Teaching Beliefs | | | 0.007 | (0.021) |
| Level-2 (Session) | | | | |
| Trainer 3 | | | -0.237 | (0.203) |
| Trainer 4 | | | -0.037 | (0.218) |
| Trainer 5 | | | -0.317 * | (0.146) |
| Trainer 6 | | | -0.278 | (0.176) |
| Trainer 7 | | | -0.354 * | (0.153) |
| Trainer 8 | | | -0.128 | (0.168) |
| Trainer 9 | | | -0.378 † | (0.215) |
| Trainer 10 | | | -0.413 † | (0.213) |
| Trainer 11 | | | -0.326 * | (0.165) |
| Trainer 12 | | | -0.366 † | (0.201) |
| Trainer 13 | | | -0.194 | (0.165) |
| Trainer 14 | | | -0.247 | (0.210) |
| Trainer 15 | | | -0.182 | (0.156) |
| Trainer 16 | | | -0.275 | (0.185) |
| Trainer 17 | | | -0.365 † | (0.207) |
| Trainer 18 | | | -0.250 | (0.183) |
| Trainer 19 | | | -0.364 * | (0.166) |
| Trainer 20 | | | -0.164 | (0.167) |
| Trainer 21 | | | -0.144 | (0.145) |
| Trainer 22 | | | -0.270 † | (0.160) |
| Trainer 23 | | | -0.138 | (0.213) |
| Trainer 24 | | | -0.181 | (0.161) |
| Trainer 25 | | | -0.329 † | (0.172) |
| Days since 11/1/08 | | | 0.000 | (0.001) |
| Number of raters | | | -0.009 | (0.006) |
| Average Ideas About Children | | | 0.002 | (0.110) |
| Average Intentional Teaching Beliefs | | | 0.050 | (0.088) |
| | Random effects | | | |
| Level-2 Intercept (uj) | 0.010 ** | (0.004) | 0.014 * | (0.005) |
| Level-1 Residual (rij) | 0.111 *** | (0.006) | 0.109 *** | (0.006) |

*Note*. Standard errors are in parentheses.

†*p* <.10. **p* <.05. ***p* <.01. ****p* <.001.

Table 5d

*Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Predictors of Absolute Distance, Survey Sample*

| | Teacher Sensitivity | | | |
|---|---|---|---|---|
| Parameter | Model 1 | | Model 2 | |
| | Fixed effects | | | |
| Intercept | 0.442 *** | (0.016) | 0.800 † | (0.466) |
| Level-1 (Rater) | | | | |
| Black | | | 0.071 | (0.044) |
| Latino | | | 0.013 | (0.056) |
| Other race/ethnicity | | | 0.024 | (0.049) |
| MA or higher level of education | | | 0.051 † | (0.030) |
| 3+ Courses in early childhood development | | | 0.007 | (0.032) |
| Num. years supervisory experience | | | 0.002 | (0.002) |
| 6+ hours per week observing | | | -0.034 | (0.029) |
| Used ECERS-R in past year | | | -0.002 | (0.030) |
| Ideas About Children | | | -0.044 | (0.033) |
| Intentional Teaching Beliefs | | | -0.029 | (0.023) |
| Level-2 (Session) | | | | |
| Trainer 3 | | | -0.166 | (0.179) |
| Trainer 4 | | | 0.176 | (0.199) |
| Trainer 5 | | | -0.054 | (0.135) |
| Trainer 6 | | | -0.188 | (0.165) |
| Trainer 7 | | | -0.030 | (0.141) |
| Trainer 8 | | | 0.112 | (0.154) |
| Trainer 9 | | | 0.044 | (0.193) |
| Trainer 10 | | | -0.052 | (0.197) |
| Trainer 11 | | | -0.043 | (0.153) |
| Trainer 12 | | | -0.078 | (0.178) |
| Trainer 13 | | | -0.015 | (0.150) |
| Trainer 14 | | | -0.074 | (0.186) |
| Trainer 15 | | | -0.025 | (0.144) |
| Trainer 16 | | | -0.010 | (0.169) |
| Trainer 17 | | | -0.047 | (0.184) |
| Trainer 18 | | | -0.003 | (0.167) |
| Trainer 19 | | | -0.014 | (0.152) |
| Trainer 20 | | | 0.076 | (0.153) |
| Trainer 21 | | | 0.047 | (0.134) |
| Trainer 22 | | | -0.079 | (0.147) |
| Trainer 23 | | | -0.052 | (0.188) |
| Trainer 24 | | | -0.047 | (0.146) |
| Trainer 25 | | | 0.075 | (0.157) |
| Days since 11/1/08 | | | 0.000 | (0.001) |
| Number of raters | | | -0.002 | (0.006) |
| Average Ideas About Children | | | -0.007 | (0.104) |
| Average Intentional Teaching Beliefs | | | -0.044 | (0.084) |
| | Random effects | | | |
| Level-2 Intercept (uj) | 0.004 | (0.003) | 0.006 | (0.005) |
| Level-1 Residual (rij) | 0.130 *** | (0.007) | 0.129 *** | (0.007) |

*Note*. Standard errors are in parentheses.

†*p* <.10. **p* <.05. ***p* <.01. ****p* <.001.

Table 5e

*Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Predictors of Absolute Distance, Survey Sample*

| | Regard for Student Perspectives | | | | | |
|---|---|---|---|---|---|---|
| Parameter | Model 1 | | | Model 2 | | |
| | Fixed effects | | | | | |
| Intercept | 0.595 | *** | (0.022) | -0.164 | | (0.536) |
| Level-1 (Rater) | | | | | | |
| Black | | | | 0.089 | † | (0.050) |
| Latino | | | | 0.117 | † | (0.064) |
| Other race/ethnicity | | | | 0.036 | | (0.056) |
| MA or higher level of education | | | | 0.003 | | (0.034) |
| 3+ Courses in early childhood development | | | | -0.023 | | (0.037) |
| Num. years supervisory experience | | | | -0.001 | | (0.002) |
| 6+ hours per week observing | | | | -0.036 | | (0.034) |
| Used ECERS-R in past year | | | | 0.035 | | (0.035) |
| Ideas About Children | | | | 0.042 | | (0.037) |
| Intentional Teaching Beliefs | | | | -0.100 | *** | (0.026) |
| Level-2 (Session) | | | | | | |
| Trainer 3 | | | | -0.001 | | (0.207) |
| Trainer 4 | | | | -0.093 | | (0.230) |
| Trainer 5 | | | | -0.011 | | (0.155) |
| Trainer 6 | | | | -0.118 | | (0.190) |
| Trainer 7 | | | | 0.108 | | (0.163) |
| Trainer 8 | | | | -0.098 | | (0.178) |
| Trainer 9 | | | | -0.326 | | (0.223) |
| Trainer 10 | | | | -0.164 | | (0.227) |
| Trainer 11 | | | | -0.222 | | (0.175) |
| Trainer 12 | | | | 0.111 | | (0.206) |
| Trainer 13 | | | | 0.027 | | (0.173) |
| Trainer 14 | | | | 0.013 | | (0.215) |
| Trainer 15 | | | | 0.006 | | (0.166) |
| Trainer 16 | | | | -0.016 | | (0.194) |
| Trainer 17 | | | | -0.134 | | (0.213) |
| Trainer 18 | | | | -0.008 | | (0.193) |
| Trainer 19 | | | | -0.137 | | (0.175) |
| Trainer 20 | | | | -0.012 | | (0.176) |
| Trainer 21 | | | | -0.146 | | (0.154) |
| Trainer 22 | | | | -0.028 | | (0.169) |
| Trainer 23 | | | | -0.240 | | (0.217) |
| Trainer 24 | | | | -0.056 | | (0.168) |
| Trainer 25 | | | | -0.185 | | (0.181) |
| Days since 11/1/08 | | | | -0.001 | | (0.001) |
| Number of raters | | | | -0.002 | | (0.007) |
| Average Ideas About Children | | | | 0.302 | * | (0.120) |
| Average Intentional Teaching Beliefs | | | | 0.152 | | (0.096) |
| | Random effects | | | | | |
| Level-2 Intercept (uj) | 0.018 | ** | (0.006) | 0.009 | | (0.006) |
| Level-1 Residual (rij) | 0.178 | *** | (0.010) | 0.169 | *** | (0.010) |

*Note*. Standard errors are in parentheses.
†*p* <.10. **p* <.05. ***p* <.01. ****p* <.001.

Table 5f

*Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Predictors of Absolute Distance, Survey Sample*

| | Behavior Management | | | |
|---|---|---|---|---|
| Parameter | Model 1 | | Model 2 | |
| | Fixed effects | | | |
| Intercept | 0.488 *** | (0.017) | 0.424 | (0.515) |
| Level-1 (Rater) | | | | |
| Black | | | 0.069 | (0.045) |
| Latino | | | -0.063 | (0.058) |
| Other race/ethnicity | | | -0.009 | (0.050) |
| MA or higher level of education | | | -0.018 | (0.030) |
| 3+ Courses in early childhood development | | | 0.003 | (0.034) |
| Num. years supervisory experience | | | 0.002 | (0.002) |
| 6+ hours per week observing | | | 0.003 | (0.030) |
| Used ECERS-R in past year | | | -0.039 | (0.031) |
| Ideas About Children | | | 0.007 | (0.033) |
| Intentional Teaching Beliefs | | | -0.016 | (0.023) |
| Level-2 (Session) | | | | |
| Trainer 3 | | | 0.091 | (0.205) |
| Trainer 4 | | | 0.016 | (0.223) |
| Trainer 5 | | | -0.031 | (0.150) |
| Trainer 6 | | | -0.044 | (0.182) |
| Trainer 7 | | | 0.125 | (0.158) |
| Trainer 8 | | | -0.012 | (0.173) |
| Trainer 9 | | | -0.008 | (0.220) |
| Trainer 10 | | | 0.009 | (0.220) |
| Trainer 11 | | | 0.111 | (0.170) |
| Trainer 12 | | | -0.047 | (0.204) |
| Trainer 13 | | | 0.096 | (0.169) |
| Trainer 14 | | | -0.014 | (0.213) |
| Trainer 15 | | | -0.039 | (0.161) |
| Trainer 16 | | | 0.068 | (0.190) |
| Trainer 17 | | | 0.115 | (0.211) |
| Trainer 18 | | | 0.204 | (0.188) |
| Trainer 19 | | | 0.075 | (0.170) |
| Trainer 20 | | | 0.085 | (0.172) |
| Trainer 21 | | | 0.103 | (0.150) |
| Trainer 22 | | | 0.018 | (0.164) |
| Trainer 23 | | | 0.164 | (0.216) |
| Trainer 24 | | | 0.156 | (0.165) |
| Trainer 25 | | | 0.162 | (0.176) |
| Days since 11/1/08 | | | 0.000 | (0.001) |
| Number of raters | | | -0.011 | (0.007) |
| Average Ideas About Children | | | -0.046 | (0.115) |
| Average Intentional Teaching Beliefs | | | 0.050 | (0.092) |
| | Random effects | | | |
| Level-2 Intercept (uj) | 0.007 | (0.004) | 0.012 * | (0.006) |
| Level-1 Residual (rij) | 0.133 *** | (0.008) | 0.132 *** | (0.008) |

*Note*. Standard errors are in parentheses.

†*p* <.10. **p* <.05. ***p* <.01. ****p* <.001.

Table 5g

*Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Predictors of Absolute Distance, Survey Sample*

| | Productivity | | | |
|---|---|---|---|---|
| Parameter | Model 1 | | Model 2 | |
| | Fixed effects | | | |
| Intercept | 0.426 *** | (0.014) | 0.772 * | (0.392) |
| Level-1 (Rater) | | | | |
| Black | | | 0.039 | (0.042) |
| Latino | | | 0.086 | (0.053) |
| Other race/ethnicity | | | 0.078 † | (0.047) |
| MA or higher level of education | | | 0.026 | (0.029) |
| 3+ Courses in early childhood development | | | 0.032 | (0.032) |
| Num. years supervisory experience | | | -0.001 | (0.002) |
| 6+ hours per week observing | | | -0.005 | (0.029) |
| Used ECERS-R in past year | | | -0.026 | (0.029) |
| Ideas About Children | | | 0.033 | (0.032) |
| Intentional Teaching Beliefs | | | -0.030 | (0.023) |
| Level-2 (Session) | | | | |
| Trainer 3 | | | -0.208 | (0.138) |
| Trainer 4 | | | -0.029 | (0.163) |
| Trainer 5 | | | -0.234 * | (0.111) |
| Trainer 6 | | | -0.307 * | (0.141) |
| Trainer 7 | | | -0.181 | (0.117) |
| Trainer 8 | | | -0.173 | (0.127) |
| Trainer 9 | | | -0.047 | (0.154) |
| Trainer 10 | | | -0.261 | (0.164) |
| Trainer 11 | | | -0.165 | (0.126) |
| Trainer 12 | | | -0.244 † | (0.138) |
| Trainer 13 | | | -0.160 | (0.120) |
| Trainer 14 | | | -0.121 | (0.145) |
| Trainer 15 | | | -0.178 | (0.120) |
| Trainer 16 | | | -0.125 | (0.137) |
| Trainer 17 | | | 0.062 | (0.144) |
| Trainer 18 | | | -0.126 | (0.137) |
| Trainer 19 | | | -0.255 * | (0.124) |
| Trainer 20 | | | -0.149 | (0.124) |
| Trainer 21 | | | -0.113 | (0.111) |
| Trainer 22 | | | -0.152 | (0.120) |
| Trainer 23 | | | -0.028 | (0.146) |
| Trainer 24 | | | -0.078 | (0.118) |
| Trainer 25 | | | -0.032 | (0.128) |
| Days since 11/1/08 | | | 0.000 | (0.000) |
| Number of raters | | | 0.000 | (0.005) |
| Average Ideas About Children | | | -0.053 | (0.089) |
| Average Intentional Teaching Beliefs | | | -0.025 | (0.072) |
| | Random effects | | | |
| Level-2 Intercept (uj) | 0.001 | (0.002) | -0.001 | (0.003) |
| Level-1 Residual (rij) | 0.125 *** | (0.007) | 0.125 *** | (0.007) |

*Note*. Standard errors are in parentheses.

†*p* <.10. **p* <.05. ***p* <.01. ****p* <.001.

Table 5h

*Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Predictors of Absolute Distance, Survey Sample*

| | Instructional Learning Formats | | | | | |
|---|---|---|---|---|---|---|
| Parameter | Model 1 | | | Model 2 | | |
| | Fixed effects | | | | | |
| Intercept | 0.465 | *** | (0.019) | 1.019 | * | (0.469) |
| Level-1 (Rater) | | | | | | |
| Black | | | | 0.029 | | (0.045) |
| Latino | | | | -0.002 | | (0.058) |
| Other race/ethnicity | | | | -0.003 | | (0.051) |
| MA or higher level of education | | | | 0.026 | | (0.031) |
| 3+ Courses in early childhood development | | | | 0.032 | | (0.034) |
| Num. years supervisory experience | | | | 0.002 | | (0.002) |
| 6+ hours per week observing | | | | 0.016 | | (0.031) |
| Used ECERS-R in past year | | | | 0.039 | | (0.032) |
| Ideas About Children | | | | -0.018 | | (0.034) |
| Intentional Teaching Beliefs | | | | -0.022 | | (0.024) |
| Level-2 (Session) | | | | | | |
| Trainer 3 | | | | -0.235 | | (0.177) |
| Trainer 4 | | | | -0.099 | | (0.199) |
| Trainer 5 | | | | -0.264 | † | (0.135) |
| Trainer 6 | | | | -0.233 | | (0.167) |
| Trainer 7 | | | | -0.033 | | (0.142) |
| Trainer 8 | | | | -0.238 | | (0.155) |
| Trainer 9 | | | | -0.155 | | (0.192) |
| Trainer 10 | | | | -0.325 | | (0.198) |
| Trainer 11 | | | | -0.181 | | (0.153) |
| Trainer 12 | | | | -0.254 | | (0.176) |
| Trainer 13 | | | | -0.069 | | (0.149) |
| Trainer 14 | | | | 0.033 | | (0.184) |
| Trainer 15 | | | | -0.133 | | (0.145) |
| Trainer 16 | | | | -0.188 | | (0.168) |
| Trainer 17 | | | | -0.343 | † | (0.182) |
| Trainer 18 | | | | 0.000 | | (0.167) |
| Trainer 19 | | | | -0.324 | * | (0.152) |
| Trainer 20 | | | | -0.224 | | (0.153) |
| Trainer 21 | | | | -0.150 | | (0.135) |
| Trainer 22 | | | | -0.256 | † | (0.147) |
| Trainer 23 | | | | -0.184 | | (0.186) |
| Trainer 24 | | | | -0.210 | | (0.146) |
| Trainer 25 | | | | -0.175 | | (0.157) |
| Days since 11/1/08 | | | | -0.001 | ** | (0.001) |
| Number of raters | | | | 0.001 | | (0.006) |
| Average Ideas About Children | | | | 0.098 | | (0.106) |
| Average Intentional Teaching Beliefs | | | | -0.090 | | (0.085) |
| | Random effects | | | | | |
| Level-2 Intercept (uj) | 0.011 | ** | (0.004) | 0.005 | | (0.004) |
| Level-1 Residual (rij) | 0.141 | *** | (0.008) | 0.143 | *** | (0.008) |

*Note*. Standard errors are in parentheses.

†*p* <.10. **p* <.05. ***p* <.01. ****p* <.001.

Table 5i

*Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Predictors of Absolute Distance, Survey Sample*

| | Concept Development | | | |
|---|---|---|---|---|
| Parameter | Model 1 | | Model 2 | |
| | Fixed effects | | | |
| Intercept | 0.980 *** | (0.034) | 2.143 ** | (0.816) |
| Level-1 (Rater) | | | | |
| Black | | | 0.193 ** | (0.071) |
| Latino | | | 0.342 *** | (0.091) |
| Other race/ethnicity | | | 0.164 * | (0.080) |
| MA or higher level of education | | | -0.129 ** | (0.049) |
| 3+ Courses in early childhood development | | | -0.045 | (0.053) |
| Num. years supervisory experience | | | 0.001 | (0.003) |
| 6+ hours per week observing | | | 0.039 | (0.048) |
| Used ECERS-R in past year | | | -0.018 | (0.049) |
| Ideas About Children | | | 0.119 * | (0.053) |
| Intentional Teaching Beliefs | | | -0.110 ** | (0.037) |
| Level-2 (Session) | | | | |
| Trainer 3 | | | -0.205 | (0.325) |
| Trainer 4 | | | 0.075 | (0.354) |
| Trainer 5 | | | 0.082 | (0.238) |
| Trainer 6 | | | -0.110 | (0.289) |
| Trainer 7 | | | -0.240 | (0.250) |
| Trainer 8 | | | -0.380 | (0.273) |
| Trainer 9 | | | -0.593 † | (0.348) |
| Trainer 10 | | | -0.097 | (0.348) |
| Trainer 11 | | | -0.005 | (0.269) |
| Trainer 12 | | | -0.137 | (0.323) |
| Trainer 13 | | | -0.246 | (0.267) |
| Trainer 14 | | | 0.545 | (0.337) |
| Trainer 15 | | | 0.111 | (0.255) |
| Trainer 16 | | | -0.160 | (0.301) |
| Trainer 17 | | | -0.149 | (0.333) |
| Trainer 18 | | | -0.312 | (0.298) |
| Trainer 19 | | | -0.488 † | (0.270) |
| Trainer 20 | | | 0.038 | (0.272) |
| Trainer 21 | | | -0.148 | (0.237) |
| Trainer 22 | | | 0.168 | (0.260) |
| Trainer 23 | | | -0.147 | (0.342) |
| Trainer 24 | | | -0.140 | (0.260) |
| Trainer 25 | | | -0.100 | (0.279) |
| Days since 11/1/08 | | | -0.001 | (0.001) |
| Number of raters | | | -0.004 | (0.010) |
| Average Ideas About Children | | | -0.122 | (0.182) |
| Average Intentional Teaching Beliefs | | | -0.129 | (0.146) |
| | Random effects | | | |
| Level-2 Intercept (uj) | 0.049 ** | (0.015) | 0.030 * | (0.015) |
| Level-1 Residual (rij) | 0.375 *** | (0.021) | 0.336 *** | (0.019) |

*Note.* Standard errors are in parentheses.

†*p* <.10. **p* <.05. ***p* <.01. ****p* <.001.

Table 5j

*Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Predictors of Absolute Distance, Survey Sample*

| | Quality of Feedback | | | | | |
|---|---|---|---|---|---|---|
| Parameter | Model 1 | | | Model 2 | | |
| | Fixed effects | | | | | |
| Intercept | 0.723 | *** | (0.032) | 0.934 | | (0.745) |
| Level-1 (Rater) | | | | | | |
| Black | | | | 0.147 | * | (0.063) |
| Latino | | | | 0.242 | ** | (0.080) |
| Other race/ethnicity | | | | 0.058 | | (0.071) |
| MA or higher level of education | | | | -0.073 | † | (0.043) |
| 3+ Courses in early childhood development | | | | -0.077 | | (0.047) |
| Num. years supervisory experience | | | | 0.002 | | (0.003) |
| 6+ hours per week observing | | | | 0.056 | | (0.042) |
| Used ECERS-R in past year | | | | -0.032 | | (0.044) |
| Ideas About Children | | | | 0.040 | | (0.047) |
| Intentional Teaching Beliefs | | | | -0.096 | ** | (0.033) |
| Level-2 (Session) | | | | | | |
| Trainer 3 | | | | -0.021 | | (0.300) |
| Trainer 4 | | | | 0.262 | | (0.325) |
| Trainer 5 | | | | 0.216 | | (0.218) |
| Trainer 6 | | | | -0.010 | | (0.263) |
| Trainer 7 | | | | -0.168 | | (0.228) |
| Trainer 8 | | | | -0.320 | | (0.250) |
| Trainer 9 | | | | -0.453 | | (0.320) |
| Trainer 10 | | | | -0.373 | | (0.318) |
| Trainer 11 | | | | 0.013 | | (0.247) |
| Trainer 12 | | | | -0.198 | | (0.298) |
| Trainer 13 | | | | -0.143 | | (0.245) |
| Trainer 14 | | | | 0.411 | | (0.311) |
| Trainer 15 | | | | 0.230 | | (0.233) |
| Trainer 16 | | | | -0.190 | | (0.276) |
| Trainer 17 | | | | 0.037 | | (0.307) |
| Trainer 18 | | | | -0.023 | | (0.273) |
| Trainer 19 | | | | -0.308 | | (0.247) |
| Trainer 20 | | | | 0.163 | | (0.249) |
| Trainer 21 | | | | -0.032 | | (0.217) |
| Trainer 22 | | | | 0.148 | | (0.238) |
| Trainer 23 | | | | -0.011 | | (0.316) |
| Trainer 24 | | | | 0.046 | | (0.239) |
| Trainer 25 | | | | -0.061 | | (0.256) |
| Days since 11/1/08 | | | | -0.001 | | (0.001) |
| Number of raters | | | | -0.007 | | (0.009) |
| Average Ideas About Children | | | | 0.100 | | (0.165) |
| Average Intentional Teaching Beliefs | | | | 0.013 | | (0.132) |
| | Random effects | | | | | |
| Level-2 Intercept (uj) | 0.047 | *** | (0.013) | 0.028 | * | (0.406) |
| Level-1 Residual (rij) | 0.277 | *** | (0.016) | 0.261 | *** | (0.056) |

*Note*. Standard errors are in parentheses.
†*p* <.10. **p* <.05. ***p* <.01. ****p* <.001.

Table 5k

*Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Predictors of Absolute Distance, Survey Sample*

| | Language Modeling | | | |
|---|---|---|---|---|
| Parameter | Model 1 | | Model 2 | |
| | Fixed effects | | | |
| Intercept | 0.938 *** | (0.037) | 0.854 | (0.834) |
| Level-1 (Rater) | | | | |
| Black | | | 0.185 * | (0.072) |
| Latino | | | 0.243 ** | (0.092) |
| Other race/ethnicity | | | 0.160 * | (0.081) |
| MA or higher level of education | | | -0.081 † | (0.049) |
| 3+ Courses in early childhood development | | | -0.054 | (0.054) |
| Num. years supervisory experience | | | 0.000 | (0.003) |
| 6+ hours per week observing | | | 0.108 * | (0.049) |
| Used ECERS-R in past year | | | -0.072 | (0.051) |
| Ideas About Children | | | 0.108 * | (0.054) |
| Intentional Teaching Beliefs | | | -0.101 ** | (0.038) |
| Level-2 (Session) | | | | |
| Trainer 3 | | | -0.075 | (0.333) |
| Trainer 4 | | | 0.414 | (0.362) |
| Trainer 5 | | | 0.251 | (0.243) |
| Trainer 6 | | | 0.012 | (0.295) |
| Trainer 7 | | | -0.152 | (0.255) |
| Trainer 8 | | | -0.229 | (0.280) |
| Trainer 9 | | | -0.444 | (0.356) |
| Trainer 10 | | | -0.240 | (0.356) |
| Trainer 11 | | | 0.088 | (0.276) |
| Trainer 12 | | | -0.075 | (0.331) |
| Trainer 13 | | | -0.137 | (0.273) |
| Trainer 14 | | | 0.669 † | (0.345) |
| Trainer 15 | | | 0.232 | (0.260) |
| Trainer 16 | | | -0.013 | (0.308) |
| Trainer 17 | | | 0.278 | (0.341) |
| Trainer 18 | | | -0.111 | (0.305) |
| Trainer 19 | | | -0.408 | (0.276) |
| Trainer 20 | | | 0.229 | (0.278) |
| Trainer 21 | | | -0.153 | (0.242) |
| Trainer 22 | | | 0.271 | (0.266) |
| Trainer 23 | | | -0.093 | (0.350) |
| Trainer 24 | | | -0.089 | (0.267) |
| Trainer 25 | | | 0.114 | (0.285) |
| Days since 11/1/08 | | | -0.001 | (0.001) |
| Number of raters | | | -0.013 | (0.011) |
| Average Ideas About Children | | | -0.019 | (0.186) |
| Average Intentional Teaching Beliefs | | | 0.110 | (0.149) |
| | Random effects | | | |
| Level-2 Intercept (uj) | 0.062 *** | (0.017) | 0.032 * | (0.015) |
| Level-1 Residual (rij) | 0.374 *** | (0.021) | 0.345 *** | (0.020) |

*Note*. Standard errors are in parentheses.

†*p* <.10. **p* <.05. ***p* <.01. ****p* <.001.